
BIG DATA
大数据
引爆新的价值点
孙 静 ◎ 主编 栾大龙 ◎ 主审
清華大學出版社

大数据：引爆新的价值点

孙　静　等 编著
栾大龙　主审

清华大学出版社
北 京

内容简介

大数据是“互联网＋”浪潮下的重要产物，也是推进“互联网＋”战略的关键技术。本书分为4篇，共10章，搜集了来自互联网企业、运营商、旅游、交通、电力、税务多个领域的真实案例，通过理论概念和应用案例相结合的方式逐步展开。从认识大数据的概念及其特征出发；定义大数据的经济资源属性，所涉及的隐私博弈和开放共享成为大数据发展的关键瓶颈；从实际应用中发现大数据的价值，不仅存在于产业经济中，还存在于社会政务中。

图书在版编目(CIP)数据

大数据：引爆新的价值点/孙静主编. —北京：清华大学出版社，2018
ISBN 978-7-302-48458-5

Ⅰ. ①大…　Ⅱ. ①孙…　Ⅲ. ①数据管理　Ⅳ. ①TP274

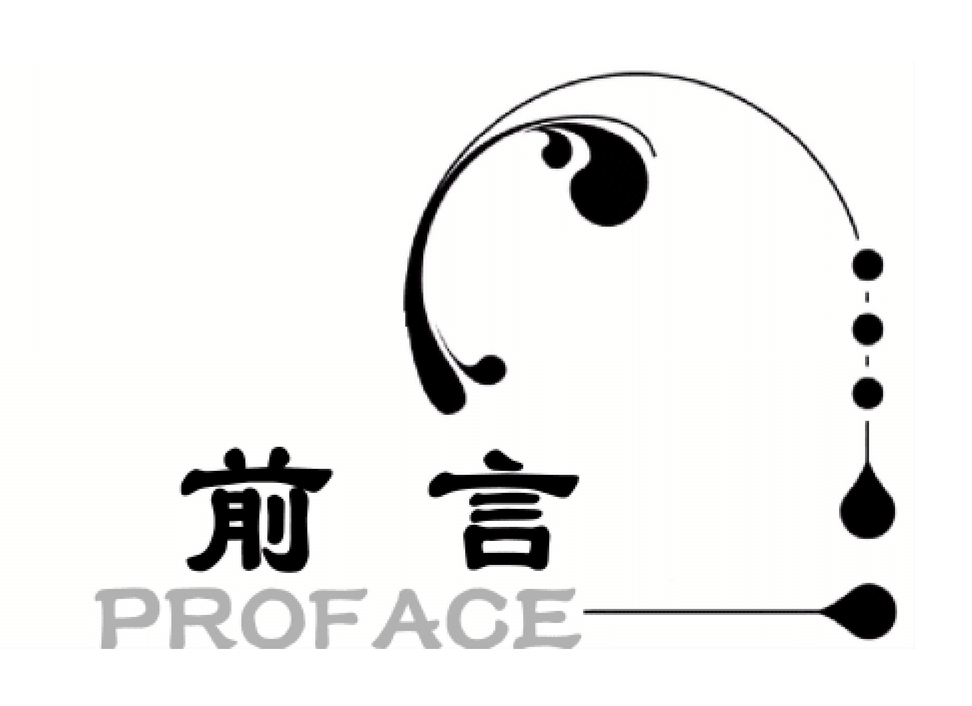
中国版本图书馆CIP数据核字(2017)第227150号

责任编辑： 贾　斌　薛　阳
封面设计： 刘　键
责任校对： 梁　毅
责任印制： 杨　艳

出版发行： 清华大学出版社
网　　址： http://www.tup.com.cn，http://www.wqbook.com
地　　址： 北京清华大学学研大厦A座　　**邮　　编：** 100084
社 总 机： 010-62770175　　**邮　　购：** 010-62786544
投稿与读者服务： 010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈： 010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载： http://www.tup.com.cn，010-62795954
印 装 者： 北京嘉实印刷有限公司
经　　销： 全国新华书店
开　　本： 185mm×260mm　　**印　　张：** 11　　**字　　数：** 262千字
版　　次： 2018年10月第1版　　**印　　次：** 2018年10月第1次印刷
印　　数： 1～2000
定　　价： 45.00元

产品编号：072282-01

编辑委员会名单



1980 年，未来学家阿尔文·托夫勒将大数据称作“第三次浪潮的华彩乐章”，与大数据相关的概念、技术、应用开始进入人们的视野，人们开始重新认识在互联网时代各类信息与行为数据所具有的更深层的意义和价值。2016 年，在中国杭州召开的二十国集团（G20）领导人第十一次峰会，大数据流动的便利化，已被认为是未来国家经济发展最重要的动力。

在微信、QQ、陌陌、云集、淘宝、京东、手机银行、高德地图、滴滴、携程、网盘、云端、手游等各类应用占满我们的手机屏幕的时候，每个人都身处大数据的旋涡之中。谁在产生数据？谁在获取数据？谁在交易数据？谁在分析数据？谁又在从数据中获取价值呢？

大数据是新兴的概念，却不是新的事物，因为数据是对客观事件进行观察或记录的结果，是我们生活或生产过程中每一个选择、决策、交流、发布、分享等行为的产物，可以说数据在人类之初就有，直到现代数字信息技术成熟，当 TB 甚至 EB 量级的数据可以被采集、存储和处理时，才有了大数据。

作为未来经济发展的“新型石油”，数据的产生是广泛的、随机的、多源的、离散的。从事大数据工作就是要发现、采集、分析、研究 TB 甚至 EB 量级的数据，从无序的数据中找到微妙的关联和规律，例如“啤酒与尿不湿”，从而对当前决策进行优化，对未来趋势进行预测。

本书编写的初衷是在与很多大数据从业者交流中，发现大数据的应用已比人们认识的更广泛，发展得更迅速，尽管业内生态仍需要完善，数据孤岛依然存在，相关政策和标准还有待制定，但是总能听到两个声音：“我们有好的实践案例想与大家分享，听取更多的建议。”“我们想看到更多、更新、更详细的案例，想将我们的数据变成真的黄金。”因此，得到北京邮电大学、阿里云研究中心、电信科学技术研究院、国家信息中心、京东物流、美林数据、明略数据等单位的大数据研究人员的支持和无私的分享，将各个领域的应用案例推荐给广大的读者。期望通过本书，能够更好地帮助大家认识大数据，理解大数据，运用大数据，在数据的海洋中挖掘到更多的宝藏。

“互联网＋”时代，信息的数字化、移动应用和支付的普及化、位置和医疗等个人信息的实时获取、物联网万事万物的互联互通……新科技、新概念如同海浪般，一浪推着一浪翻滚前进，大数据就如同海中的浪花一般随波前行，且随着一浪一浪的相互推动而越发丰富绚烂。

编　者
2018 年 6 月



第一篇　大数据，新时代的代名词

第 1 章　数据时代 …… 3
1.1　大数据溯源 …… 3
1.1.1　数据起源 …… 4
1.1.2　数据存储 …… 4
1.1.3　数据计算 …… 6
1.2　初识大数据 …… 7
1.2.1　大数据的定义 …… 7
1.2.2　大数据的特征 …… 9
1.2.3　大数据与传统数据分析的区别 …… 11
1.3　大数据应用的演进趋势 …… 12
附：大数据年代记 …… 14
第 2 章　大数据关键技术 …… 17
2.1　物联网 …… 18
2.1.1　物联网的概念 …… 18
2.1.2　物联网：大数据资源的重要提供者 …… 20
2.2　移动互联网 …… 21
2.2.1　移动互联网的发展 …… 21
2.2.2　移动互联网：大数据的传输载体 …… 23
2.3　云计算 …… 24
2.3.1　云计算的优点 …… 24
2.3.2　云计算与大数据的关系 …… 25
2.4　智慧旅游的大数据采集 …… 26
2.4.1　整合内外部数据 …… 26
2.4.2　信息化平台——数据采集存储基础设施 …… 27

第二篇　大数据，一种经济资源

第 3 章　数据价值与隐私博弈 …… 31
3.1　数据的经济属性 …… 31

3.1.1　经济物品与经济资源 …… 31
3.1.2　数据信息转化为经济资源 …… 34
3.2　大数据时代的个人隐私 …… 36
3.2.1　隐私的数据化 …… 37
3.2.2　数据的商业化 …… 38
3.2.3　个性化服务的博弈 …… 39
3.3　旅游大数据应用价值 …… 40
3.3.1　数据推动旅游行业价值流动 …… 40
3.3.2　智能服务 …… 41
3.3.3　智慧管理 …… 43

第4章　大数据的开放与共享 …… 47

4.1　数据资源开放和共享 …… 47
4.1.1　打破“信息孤岛” …… 47
4.1.2　全球数据的开放与共享 …… 48
4.1.3　数据标准化 …… 49
4.2　数据构建的“知识森林” …… 51
4.2.1　平台设计 …… 52
4.2.2　实施路径 …… 53
4.2.3　案例分析 …… 56

第三篇　大数据，价值创新的土壤

第5章　大数据精准营销 …… 61

5.1　大数据营销 …… 61
5.1.1　精准营销 …… 61
5.1.2　精准广告 …… 64
5.2　实时竞价广告 …… 64
5.2.1　RTB广告投放关键技术 …… 65
5.2.2　RTB的生态圈 …… 66
5.2.3　RTB投放工作内容 …… 70
5.2.4　RTB应用场景示例 …… 71

第6章　阿里的数据王国 …… 74

6.1　“滴滴打车”助市民出行无忧 …… 75
6.1.1　典型案例 …… 76
6.1.2　案例分析 …… 77
6.1.3　“快的”的数据价值 …… 80
6.2　“聚划算”的智慧营销 …… 81

6.2.1 商家端:数据化招商 …… 81
6.2.2 消费者端:数据化导购 …… 85
6.2.3 “聚划算”的数据价值 …… 91

第 7 章 让数据告诉你“谁可信” …… 92

7.1 “区块”成“链” …… 92
7.1.1 区块的形成 …… 93
7.1.2 区块链的特征 …… 94
7.2 “芝麻信用”让信用等于财富 …… 95
7.2.1 什么是信用 …… 95
7.2.2 从“信用”到“财富” …… 96
7.2.3 信用商圈 …… 101

第 8 章 大数据“地图” …… 103

8.1 便捷交通大数据服务 …… 103
8.1.1 城市公共交通存在的问题及其现状 …… 104
8.1.2 大数据服务应用 …… 104
8.1.3 个人应用场景 …… 107
8.2 人群流动监控 …… 109
8.3 实时车流控制系统 …… 111

第四篇 大数据,推动新型政务

第 9 章 大数据时代的税务精细化管理 …… 117

9.1 大数据时代下税务工作新趋势 …… 117
9.1.1 税务数据新趋势 …… 117
9.1.2 税务业务新趋势 …… 118
9.1.3 新机遇和新挑战 …… 119
9.2 大数据技术的价值 …… 120
9.3 税务精细化管理顶层设计 …… 121
9.4 大数据税务应用整体架构 …… 122
9.4.1 总体架构 …… 122
9.4.2 汇集层 …… 122
9.4.3 数据层 …… 123
9.4.4 服务层 …… 123
9.4.5 应用层 …… 124
9.4.6 大数据安全保障体系 …… 124
9.4.7 运维保障 …… 125
9.4.8 标准化体系 …… 125

9.5 大数据税务数据应用场景 …… 125
9.5.1 优化纳税服务 …… 125
9.5.2 社会经营关系 …… 127
9.5.3 偷税漏税 …… 128
9.5.4 涉税事件追踪 …… 129
9.5.5 税源画像 …… 130
9.5.6 纳税遵从指数 …… 131
9.6 税务大数据服务价值 …… 133

第10章 大数据时代的电力服务 …… 135

10.1 电力大数据面临挑战 …… 135
10.2 电网运营大数据 …… 136
10.2.1 电网系统架构现状 …… 136
10.2.2 高性能架构设计 …… 136
10.2.3 技术选型 …… 139
10.2.4 高性能架构设计实践 …… 142
10.3 电网用户行为分析 …… 145
10.3.1 分析目标及原则 …… 145
10.3.2 用电行为分析总体架构 …… 147
10.3.3 宏观层面用电行为分析 …… 150
10.3.4 微观层面用电行为分析 …… 156

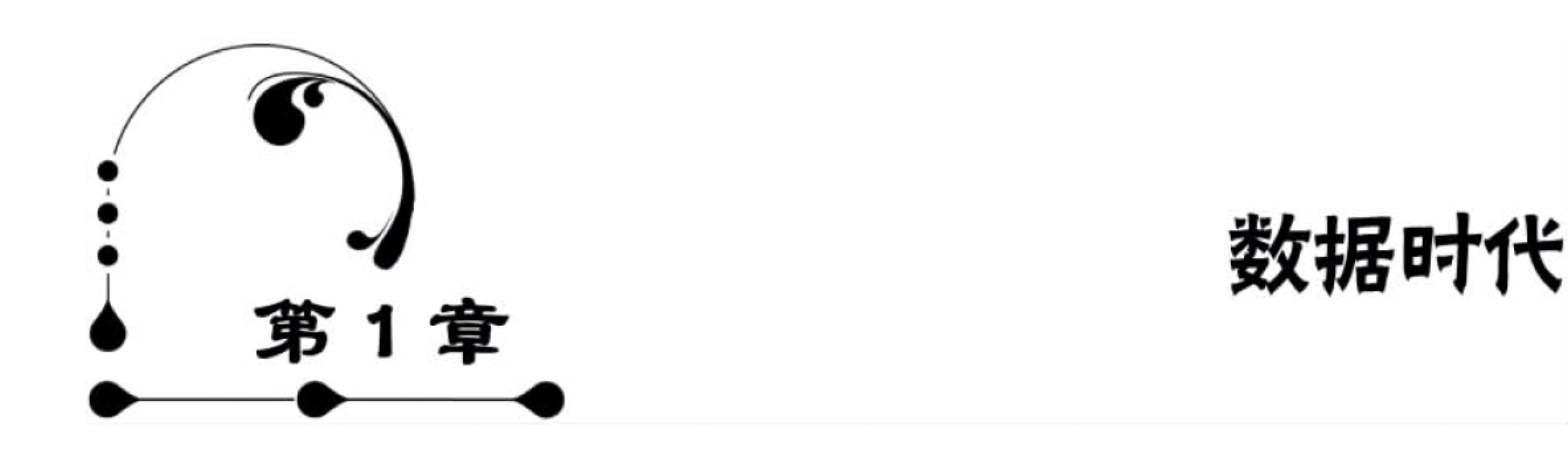
第一篇　大数据，新时代的代名词

人类历史上从未有哪个时代和今天一样产生如此海量的数据，数据的产生已经完全不受时间、地点的限制，尤其是社交网络、电子商务、移动互联网等领域的飞速发展，把人类社会带入了一个以PB（1PB＝1024TB，1TB＝1024GB）为单位的结构和非结构数据构成的网络化、数字化时代。一个大规模生产、分享和应用数据的时代正在开启。

2016年10月，在杭州举行的"互联网大数据高峰论坛"上，阿里巴巴原副总裁、《数据之巅》的作者涂子沛指出："弄潮儿向涛头立，手把红旗旗不湿。我们今天的涛头就是互联网、大数据。"

中国科学院大学经管学院教授吕本富，在《G20国家互联网发展研究报告》中指出"我们认为，未来数据的流动将是世界经济增长最重要的动力来源。如果过去是服务的便利化、贸易的便利化，以后一定是大数据流动的便利化，成为这个国家经济发展最重要的动力。"

大数据已掀起了时代的浪潮，任何人和事都需要用数据说话！



第1章 数据时代

“大数据”一词近五年在百度搜索指数中的整体趋势从2012年开始呈快速增长的态势，并在2016年5月周平均值最高达7287，如图1-1所示。

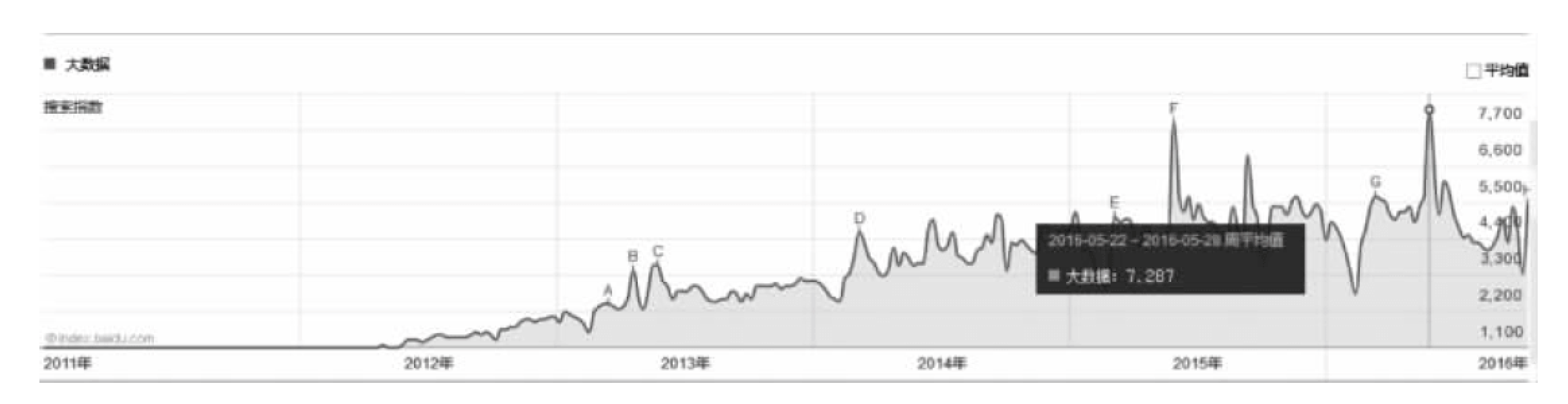


图1-1 “大数据”的百度搜索指数

（数据来源：百度指数，http://index.baidu.com/）

2014年8月，在中央电视台财经频道、综合频道、纪录频道、科教频道播出了一套10集的纪录片《互联网时代》（别名《大数据时代》），这是中国第一部，甚至也是全球电视机构第一次全面、系统、深入、客观地解析互联网的大型纪录片。这部在开播前没有密集的节目宣传，没有明星、噱头与谈资的纪录片，仅是在社交网络上的口口相传，百度搜索的指数就从0开始直线攀升至15127。2015年8月31日，国务院印发《促进大数据发展行动纲要》（国发〔2015〕50号）。指出主要任务是：加快政府数据开放共享，推动资源整合，提升治理能力；推动产业创新发展，培育新兴业态，助力经济转型；强化安全保障，提高管理水平，促进健康发展。

这些数据和现象都在说明一个客观事实、一个社会热点、一个发展趋势——大数据的发展已成为国家发展战略的重要组成部分，大数据正在或已经成为时代前进的代名词。如何认识大数据，如何应用大数据，是从IT时代走向DT时代的必要课题。

1.1 大数据溯源

早在1980年，著名未来学家阿尔文·托夫勒在其所著的《第三次浪潮》中就提出“数据就是财富”，并热情地将“大数据（Big Data）”称颂为“第三次浪潮的华彩乐章”。但是到2008年，学术界、工业界甚至于政府机构才开始密切关注大数据问题。*Nature* 杂志在2008年9月推出了名为“大数据”的封面专栏，*Science* 则在2011年推出了专刊 *Dealing with Data*，

主要围绕着科学研究中大数据的问题展开讨论，说明大数据对于科学研究的重要性①。

大数据的概念和技术不是凭空出现的，人们对于大数据的认知或许最早来自托夫勒在其所著的《第三次浪潮》，但是人类对于数据的搜集、存储可以追溯到远古时代，对于事物的数据化发展于计算机的出现。"大数据"并不是作为一个全新的事物出现的，它是基于人类发展过程中，对于数据搜集、存储、分析能力的提升而出现的一种新的思维方式，一种新的服务模型，一股推动经济社会发展新的助力。

1.1.1 数据起源

数据(data)是对客观事件进行观察或记录的结果，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合，是对客观事物的逻辑归纳，用于表示客观事物的未经加工的原始素材。它可以是数字，也可以是具有一定意义的文字、字母、数字符号的组合、图形、图像、视频、音频等，是可识别的对客观事物的属性、数量、位置及其相互关系的抽象表示符号。

大约两万年前的伊尚戈骨头(Ishango Bone，图 1-2)被认为是最早的记录数据和分析数据的工具，是旧石器时代人们采用在树枝或者骨头上刻下凹痕的方法来记录日常的交易活动或物品供应。



图 1-2　伊尚戈骨头

1991 年，计算机科学家蒂姆·伯纳斯·李宣告了我们今天所熟知的万维网的诞生。在一个网站上，他制定了世界网络的协议书，使互联网的数据联通起来，让任何人可以在任何地方进行通信。互联网时代的开启，带动了各行各业的网络化发展。人、物、机器等都可以通过一个终端接入这个不受时间、空间限制的虚拟网络中。在商业、生活、生产、农业、医疗、金融等领域网络化的过程中，带来了以几何倍数增长的数据量。

2004 年，Facebook(脸书)、Twitter、Instagram 等社交网络的相继问世迎来了开放共享的 Web 2.0 时代。网络平台不再是自上而下地由少数资源所有者控制，而是自下而上地由广大用户的智慧和力量主导。在 Web 2.0 模式下，网络用户出于对某个或某些问题的共同兴趣而聚集，这促使他们主动积极地参与问题讨论和信息分享。全球数据量预测如图 1-3 所示。

根据知名市场研究机构 IDC(International Data Corporation，国际数据公司)的研究报告表明，2011 年全球数据总量已经达到 1.8ZB，未来全球数据总量年增长率将维持在 50% 左右，到 2020 年，全球数据总量将达到 40ZB，如图 1-3 所示。

1.1.2 数据存储

人们在生产生活过程中所创造的各种数字、图像、文字、记录等需要被采集并保存下来，才能够形成数据。一个坚持 30 年，每天走一万步的人，他的个人运动数据和位置数据，在微信运动或计步 App 等出现后，同样的行为才被采集并存储成为数据。

亚历山大图书馆(公元前 300 年—公元 48 年)可能是古代最大的数据储存地了，这里

① 孟小峰，慈祥. 大数据管理：概念、技术与挑战[J]. 计算机研究与发展，2013，50(1)：146-169.

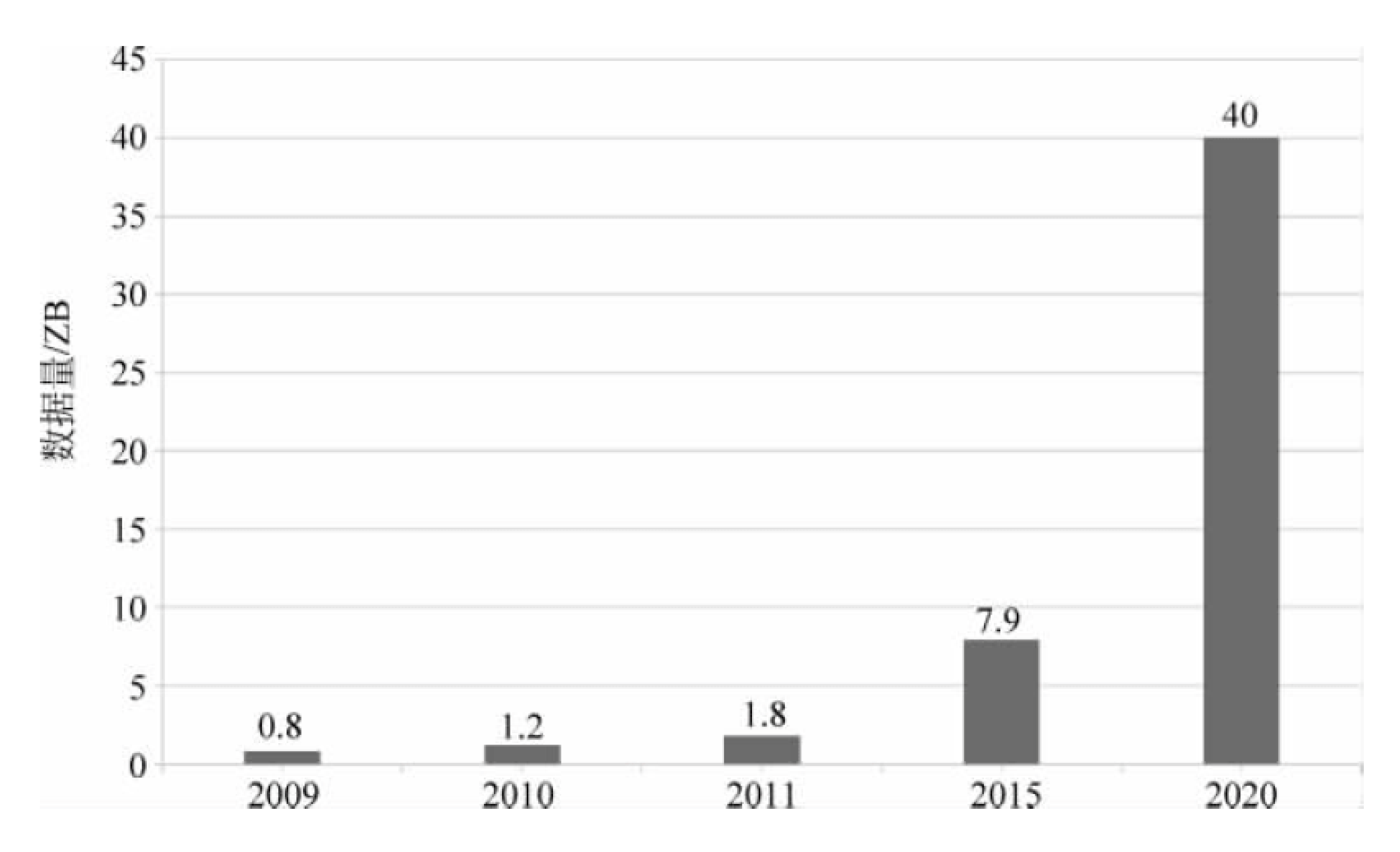


图 1-3 全球数据量预测

(数据来源：IDC)

50 万卷的藏书几乎涵盖了当时人们学习的各个领域。

1928 年，工程师波弗劳姆(Fritz Pfleumer)发明了一种用磁带来存储信息的方法。他发明的这个原理今天依然在使用，绝大部分的数据就是存储在有磁性介质的计算机硬盘上。

1965 年，英特尔(Intel)创始人之一戈登·摩尔(Gordon Moore)提出了摩尔定律，揭示了信息技术进步的速度。其内容为：当价格不变时，集成电路上可容纳的元器件的数目，约每隔 18～24 个月便会增加一倍，性能也将提升一倍。在摩尔定律的推动下，计算存储和传输数据的能力在以指数速度增长，每 GB 存储器的价格每年下降约 40%。

1965 年，美国政府计划在世界首个数据中心的磁盘上存储 7.42 亿的纳税申报单和 1.75 亿的指纹信息。1967 年，IBM 公司推出世界上第一张"软盘"，是最早的可移动数据存储介质。

2010 年印刷版《大英百科全书》，共 32 册，重达 58.5kg，然而它的全部内容，还装不满一个 4GB 的 U 盘。

历史的进程进一步证实了摩尔定律，数据存储能力的指数提升如图 1-4 所示。

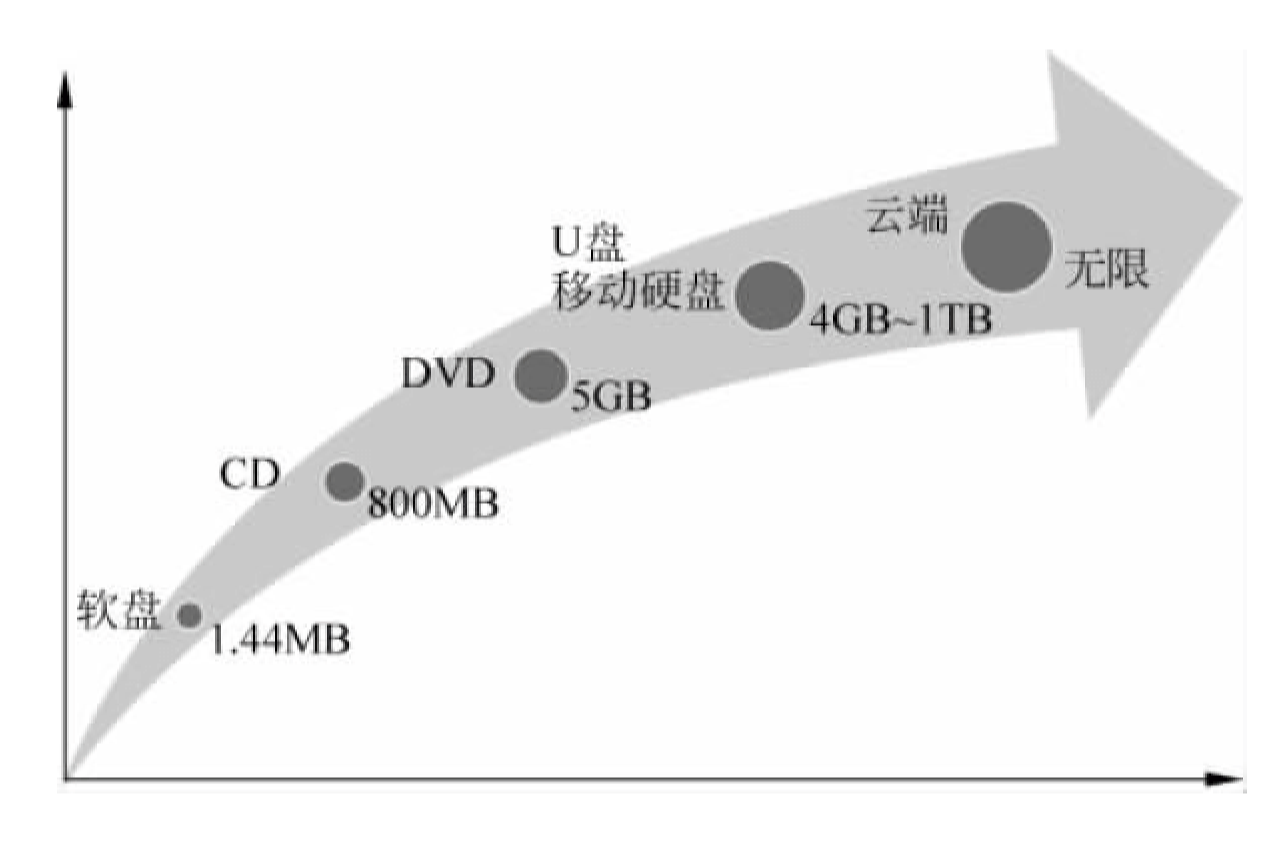


图 1-4 数据存储能力的提升

1.1.3 数据计算

数据分析就是对数据进行分析并得出有用的结论。首先不一定使用统计分析的方法；其次，不一定非要处理大量的数据，也不一定要用计算机；再次，数据分析自古就有。百度百科对于数据分析的片面认识反映了国内人们对于数据分析认识的模糊，也反映了商业利益对于正常观念的扭曲。

数据分析早在两千多年前就在使用。在战国时期的孙庞斗智中，孙膑设计蒙骗庞涓，孙膑命令部队，每日大幅减少炉灶的数量。庞涓通过观察孙膑军队的炉灶数量逐日大量减少，分析得出孙膑军队大量逃散的结论，最终上当战败。这就是数据分析。

在辽沈战役中，林彪在诸多战报中发现，在胡家窝棚附近缴获的短枪与长枪的比例比其他战斗中的高，那里缴获和击毁的小车与大车的比例比其他战斗中的高，在那里俘虏和击毙的军官与士兵的比例比其他战斗中的高。他就断定，敌人的指挥所就在这里。果不其然，敌军司令廖耀湘在胡家窝棚附近被逮个正着。这也是数据分析。

数据分析发展自古到今，已经涵盖了最朴素的数据分析，也涵盖了数据统计、数据挖掘和大数据处理的所有内容。这两个案例都说明了数据分析古已有之，且数据分析不一定要有海量数据，也不一定要用复杂度统计分析方法，只要统计数据分类(统计口径)正确；同时还说明了数据分析极其重要，更说明了数据意识和素质的重要。

当各类数据能够被采集并得以保存时，提升计算和分析数据的能力，成为实现数据价值的必要手段。

安提凯希拉(Antikythera)机器，是最早被发现的机械计算机[①]，也代表了数据分析能力从人工计算向机械计算的提升。

1663 年，约翰·葛兰特(John Graunt)在伦敦用记录下的当时肆虐欧洲的黑死病死亡人数信息，建立起了早期预警系统的理论，是第一次有记录的统计数据分析实验。1865 年，银行家亨利·福尼斯(Henry Furnese)用结构化的方式收集和分析有关竞争对手的商业活动来取得竞争优势，这被认为是第一次将数据分析用于商业目的。

1881 年，美国人口普查局聘用了一位年轻的工程师赫尔曼·何乐礼(Herman Hollerith)，他发明了著名的打孔卡片制表机，被认为是现代计算机的雏形，将原本预计需要花费 10 年时间去分析的 1880 年收集到的人口普查数据工作缩短为三个月，数据处理速度提升了近 40 倍。

1989 年，美国计算机协会(Association of Computing Machinery，ACM)数据挖掘机知识发现委员会(Special Interest Group on Knowledge Discovery and Data Mining，SIGKDD)主办了第一届数据挖掘学术年会。基于数据的采集、分类、估值、语言、相关性分组或关联规则、聚集、描述和可视化等分析方法开始深入到人们生活的方方面面。

2004 年，谷歌公开的 MapReduce 分布式并行计算技术，是新型分布式计算技术的代表。一个 MapReduce 系统由廉价的通用服务器构成，通过添加服务器节点可线性扩展系统的总处理能力(Scale Out)，在成本和可扩展性上都有巨大的优势。

2005 年，Hadoop 诞生，它是专门为存储及分析大数据的开源框架。它能够灵活管理人

---

① Antikythera mechanism[OL]. Wikipedia，https://en.wikipedia.org/wiki/Antikythera_mechanism.

们不断产生和采集的非结构化数据，例如语音、视频、文档等。以 Hadoop 为代表的分布式存储和计算技术迅猛发展，极大地提升了互联网企业数据管理能力，互联网企业对“数据废气”的挖掘利用大获成功。

2007 年，《连线》(*Wired*)杂志在文章《理论的终结：数据洪流让科学方法过时》中将“大数据”的概念引进了大众的视野①。

回顾数据的起源和发展，可以清晰地看到今天的大数据是从最朴素的数据分析、数据统计和数据挖掘一步步走过来的，数据分析为社会带来的经济价值越来越高。今天的大数据也好，数据挖掘也罢，都是在做数据分析这件事，只不过是数据的体量在提高，数据的复杂性在提高，数据处理的能力在提高以及数据处理的结果更具有创造性。

从最朴素的数据分析到大数据处理，运用数据的思路与逻辑是一致的。所有的数据分析无非是在寻找：什么是我要找的数据，我要找的数据在哪里能找到，最大(小)的数是多少，最大(小)的数据在哪里，最大(小)的可能是多少，最大(小)的可能在哪里，哪些因素最相关，相关性多大，从大到小的排序，按照时间或位置排列的升降状态等。数据分析的思路就是搜索、对比、概率计算、相关性分析、分类、排序、预测等，最后做出的结果就是预测、聚类与排序。

1.2 初识大数据

在人类社会发展的历史长河中，经济发展往往伴随着技术革命。2013 年称为“大数据元年”。目前，几乎所有世界级的互联网企业，都将业务触角延伸至大数据产业；无论社交平台逐鹿、电商价格大战还是门户网站竞争，都有它的影子。

大数据无处不在，大数据应用影响到了人们的工作、生活和学习，并将继续施加更大的影响。

1.2.1 大数据的定义

在计算机科学中，数据是指所有能输入到计算机并被计算机程序处理的符号的介质的总称，是用于输入电子计算机进行处理，具有一定意义的数字、字母、符号和模拟量等的通称②。

数据的基本计量单位是 Byte，按照 1024(2^{10})进率，依次递增为 B、KB、MB、GB、TB、PB、EB、ZB、YB、DB、NB。

1B＝8b

1KB＝1024B

1MB＝1024KB

1GB＝1024MB

1TB＝1024GB

1PB＝1024TB

① The End of Theory：The Data Deluge Makes the Scientific Method Obsolete[OL]. 2013，http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/.

② 王珊，萨师煊. 数据库系统概率(第 5 版)[D]. 北京：高等教育出版社，2014.

1EB＝1024PB

1ZB＝1024EB

1YB＝1024ZB

“大数据”一词本身就是一个比较抽象的概念，单从字面来看，“大”体现了研究或应用的量级规模是庞大的，“数据”则说明了研究或应用对象的实质。但是什么样的数据量级才可以称之为“大”呢？

传统数据库有效工作的数据规模一般为10～100TB，因此麦肯锡和IDC公司对此都有过相近的说法，10～100TB通常成为大数据的门槛。所谓大数据从数据规模上看，大概是指100TB以上的数据体量，100TB相当于现在100部最新笔记本（1TB硬盘）的最大存储总量。但是，数据计算的难度与速度还涉及数据的类型、结构与存储的复杂性，因此以100TB为基准来定义大数据的说法未必科学。

大数据和互联网都是一种通用目的的技术（General Purpose Technology），随着技术和应用的发展，其概念也在不断地演进。尽管有很多研究机构和学者给出的定义被广泛认可，但是却没有公认的、唯一的准确定义。

维克托·迈尔·舍恩伯格与肯尼斯·库克耶在他们合著的《大数据时代》一书中指出：大数据是指不用随机分析法这样的捷径，而采用所有数据的方法[①]。

大数据：样本＝全体。

因此，所谓的“大”其实也包含着“全”的含义，不是相对的量级，而是绝对的范围。

对于大数据这一概念比较被认可的定义还有以下几种。

（1）大数据，或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息。（维基百科[②]）

（2）一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模（Volume）、快速的数据流转（Velocity）、多样的数据类型（Variety）和价值密度低（Value）4大特征。（麦肯锡全球研究所）

（3）大数据是数据集或信息，它的规模、发布、位置在不同的信息孤岛上，或它的时间线要求客户部署新的架构来捕捉、存储、整合、管理和分析这些信息以便实现企业价值。（EMC公司）

（4）大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产，这些信息资产需要新型的处理方式来强化决策制定、洞察发现和处理优化。（研究机构Gartner，2012）

（5）大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。（中华人民共和国国务院，《促进大数据发展行动纲要》，2015）

---

① [英]维克托·迈尔·舍恩伯格，肯尼斯·库克耶. 大数据时代[M]. 盛阳燕，周涛，译. 浙江：浙江人民出版社，2013.

② Big data [OL]. Wikipedia，https://en.wikipedia.org/wiki/Big_data.

这些定义都强调的是大数据的海量数据规模、多样数据类型和新型处理技术的特点。Gartner将大数据定义为一种信息资产,即数据的价值不仅体现在数据本身,更可以作为市场经济中的生产要素,用于交易并创造出更大的价值。我国的《促进大数据发展行动纲要》中,将大数据作为新一代信息技术和服务业态,强调了大数据具有的创新性和服务性,是信息革命、互联网+时代引领的新型应用、新型服务、新型行业。

大数据从哪里来?我们可以把它简单地概括为以下三大类。

第一,流动数据。物质世界本身数字化产生的大数据。例如一些医疗服务类网站,将医生信息、门诊信息等现实事物数字化,形成了大量网络数据;物联网上的人、机、物交互产生了实时的行为轨迹和状态数据。

2010年,美国有1.5亿慢性病患者,如糖尿病、充血性心脏衰竭、高血压患者,他们的医疗费用占到了医疗卫生系统医疗成本的80%。远程病人监护系统对治疗慢性病患者是非常有用的。远程病人监护系统包括家用心脏监测设备、血糖仪,甚至还包括芯片药片,芯片药片被患者摄入后,实时传送数据到电子病历数据库。

第二,社交数据。用户在互联网交流过程中不断产生各式各样的行为大数据,这类数据在社交互动中越来越具有吸引力,尤其是它的营销功能。但是这些数据通常是在非结构化或半结构化形式,对于一个公司当使用和分析这些数据信息的时候,不仅要考虑数据的规模,大数据应用也是一个独特的挑战。大量移动电子终端设备的出现,更加快了互联网信息制造的速度。

2011年8月23日,美国弗吉尼亚州发生5.9级地震,纽约市民首先在Twitter上看到地震信息之后才感到震区传来的真实震感。这意味着,社交网络不但是提升人类信息传播速度的工具,也是用户随时随地记录行为、思想和情绪的平台,而这种数字化的记录就是制造数据的过程。

第三,公开来源。庞大的数据可以通过打开数据源,像美国政府的数据,CIA世界各国概况或者欧盟开放数据门户等获得。各种数据的积累、沉淀及保存产生大数据。随着科技进步,时代变化,高性能存储设备日益发展普及,使越来越多的数据得以持续保存,形成越发庞大的数据集。

国家邮政局公布2016年10月邮政行业运行情况数据:全行业业务收入完成483.5亿元,同比增长38.4%;业务总量完成695.5亿元,同比增长48.4%。其中,快递业务量完成30.3亿件,同比增长55.9%;业务收入完成376.2亿元,同比增长49.1%。

1.2.2 大数据的特征

基于全体样本的分析是"大数据"定义中对于研究对象进行界定的核心内涵,所体现出的特征也必然围绕着全体样本集合的特点。

在2001年的研究报告和相关文献中,META Group(现在的Gartner)的分析师Doug Laney将数据增长的挑战和机遇定义成三维方式,即数据总量Volume、处理速度Velocity和数据类型Variety,也就是最早用来描述大数据的"3V"模型。

随着资讯科技不断地往前推进,数据量的复杂程度愈来愈高,3V已经不足以形容新时代的大数据。2012年,包括IBM、Gartner、IDC在内的科技厂商和研究机构等纷纷提出新的论述,在3V的基础上增加了对数据"价值(Value)"的认识,发展成为4V模型。阿姆斯特

丹大学的 Yuri Demchenko 等人提出大数据还应具有可信性、真伪性、来源和信誉、有效性和可审计性的特点，即真实性（Veracity），形成了 5V 的框架，如图 1-5 所示。

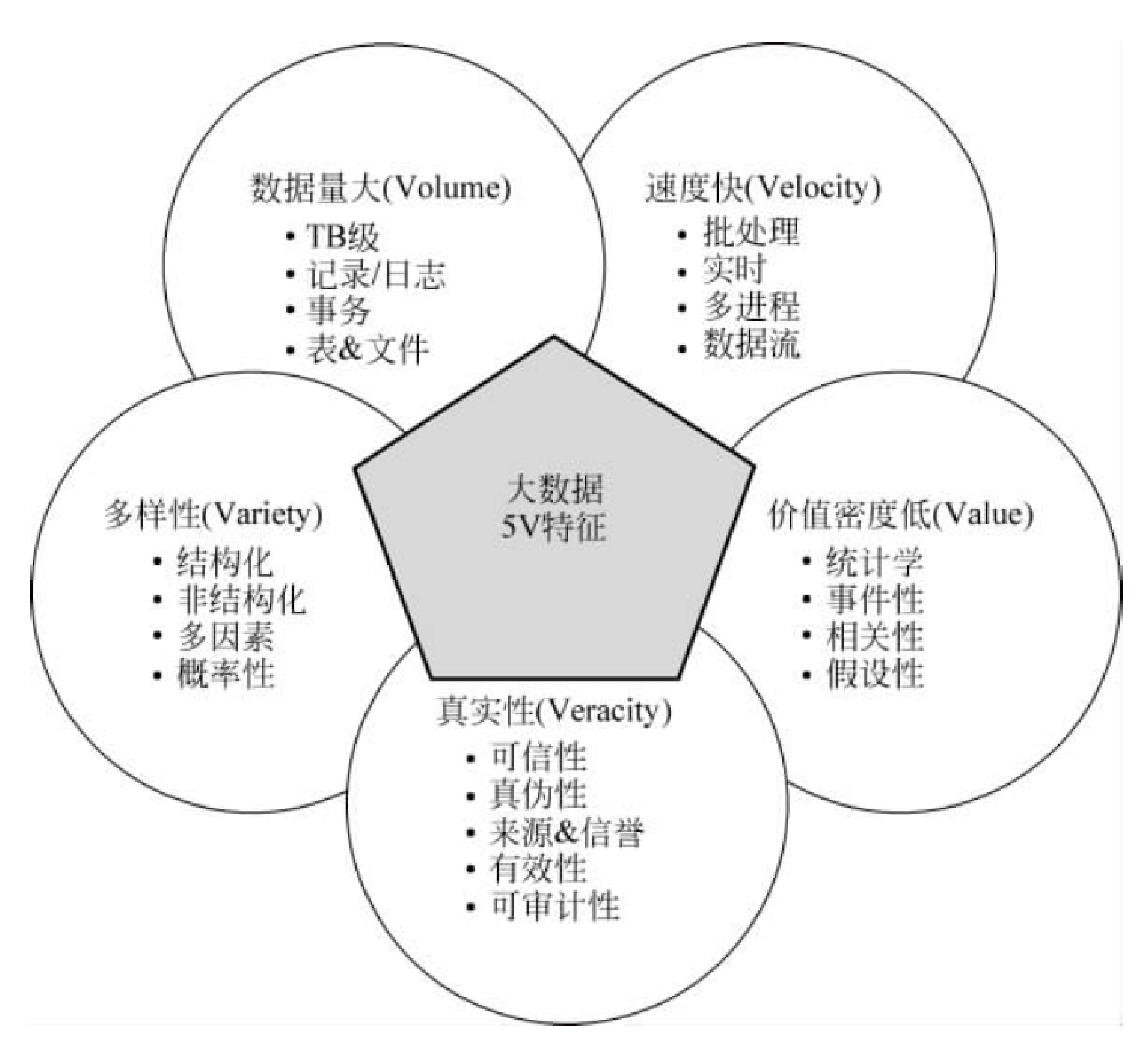


图 1-5　大数据 5V 特征

第一，数据体量巨大（Volume）。指收集和分析的数据量非常大，从 TB 级别跃升到 PB 级别，但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了 PB 级的数据量。

2006 年，个人用户每年产生的数据才刚刚迈入 TB 时代，全球一共新产生了约 180EB 的数据；在 2011 年，这个数字达到了 1.8ZB。2013 年，中国产生的数据总量超过 0.8ZB，是 2012 年的两倍，相当于 2009 年全球的数据总量。

第二，处理速度快（Velocity）。大数据需要对数据进行近实时的分析。以视频为例，连续不间断监控过程中，可能有用的数据仅有一两秒，这一点和传统的数据挖掘技术有着本质的不同。

每秒钟淘宝商城就会产生大约 178 笔订单；每分钟人们可以在 YouTube 上传 20 个小时的视频。Facebook 位于瑞典北方的“资料库”——Node Pole，有 6 个足球场那么大，每天要处理全球用户 45 亿个赞、3.5 亿张照片和 100 亿条信息。

第三，数据多样性（Variety）。大数据来自多种数据源，数据种类和格式日渐丰富，包含结构化、半结构化和非结构化等多种数据形式，如网络日志、视频、图片、地理位置信息等。非结构化数据增长率达 80%，而传统的数据样式主要以数据库和 XML 为主。

多样的数据类型涉及数字、文字、图片、语音、视频、地理位置、网络日志信息等，从数据结构来看，可分为非结构化数据、半结构化数据和结构化数据，从数据存储方案的角度还可以分为分布式存储和集中式存储，从数据质量来看，数据的完整性、可信性与可用性也大不相同。对于是否必须采用大数据运算，在数据规模和数据复杂性之间存在一定的取舍关系。一般来说，分布式存储就意味着很大的数据体量，分布存储的数据就需要用大数据技术来处

理了,传统技术已经无法使用。大数据适合于处理分布式存储的复杂数据。

第四,价值密度低(Value)。要挖掘大数据的价值就需要在几百万条数据中找到真正有借鉴意义的几条,例如每天 24 小时的视频数据中,针对某一研究或分析目标有价值的仅有几秒钟。通过分析数据得出如何抓住这条数据,就能够把握机遇并收获巨大的经济或社会价值。

第五,数据真实性(Veracity)。大数据中的内容是从真实世界采集得到的,在录入、生成、采集数据的过程中存在因为客观或人为因素产生偏差的情况。数据的真实性即代表了数据的质量,将直接影响分析和预测的准确性、真实性和有效性。大数据就是从庞大的网络数据中提取出能够解释和预测现实事件的过程。

大数据的产生和发展,是信息技术领域不同时期的多个进步交互作用的结果。在未来,智能数据可以帮助我们了解一个智能系统每时每刻发生了什么,更能够告诉人们为什么会发生。甚至还可以告诉人们接下来会发生什么,以及我们应该如何应对,智能数据将改变人们的生活方式和思维模式,提升国家或政府的服务能力,创新企业的商业模式。

1.2.3 大数据与传统数据分析的区别

大数据分析是指对大量结构化和非结构化的数据进行分析处理,从中获得新的价值,具有数据量大、数据类型多、处理速度快等特点。与传统数据分析相比,大数据分析的特点如表 1-1 所示。

表 1-1 大数据分析与传统数据分析的区别

	大数据分析	传统数据分析
数据量	TB、PB 以上	MB 至 GB
数据类型	种类繁多,包括非结构化、非线性数据	种类单一,以结构化数据为主
产生模式	现有数据才能确定模式,且模式随着数据的增长不断演化	先有模式,才会产生数据
处理能力	可对收集到的所有海量数据进行分析	对通过采样方式获得的部分数据进行分析
处理范围	将不同领域的数据组合后进行分析	单一领域内的数据分析
处理结果	数据源与分析结果间不仅是因果关系,还可基于有相关关系的数据源完成分析预测;进行机器学习等	关注数据源与分析结果间的因果关系;对数据进行查询等

数据无处不在,大数据分析不仅是对因果关系的研究,在智慧城市的建设中,还注重对相关数据的挖掘以获得对未来合理的预测分析。对农业的大数据处理,能够预测可能的病虫危害或天气变化,从而提前做好防护和灌溉规划;对交通的大数据处理,能够预测道路拥堵的状况以便提前做好疏通准备和出行计划;对医疗的大数据处理,可以预测个人身体健康状况发展及各类疾病的发生率或者就医资源的需求,以便提前做好个人的健康管理和医疗资源的调度。如果能预知下一秒可能发生什么,那么就能在当下做出最有利于下一秒的决策,始终赢在起跑线前一秒。

认识大数据,不仅是要认识数据本身,还需要处理和分析数据的模型、技术、手段等。大数据不是一个简单的实物名词,而是围绕全样本数据的一系列计算、分析,以及获得的有效

信息或智能预测。通过数据的采集、分析模型的建立、计算工具的应用，最终实现从数据到具有商业价值的信息资产的转变，这就是大数据技术。

1.3 大数据应用的演进趋势

大数据行业应用的发展，是沿袭数据分析应用而来的渐变的过程。观察大数据应用的发展演变，可以从技术强度、数据广度和应用深度三个视角切入。

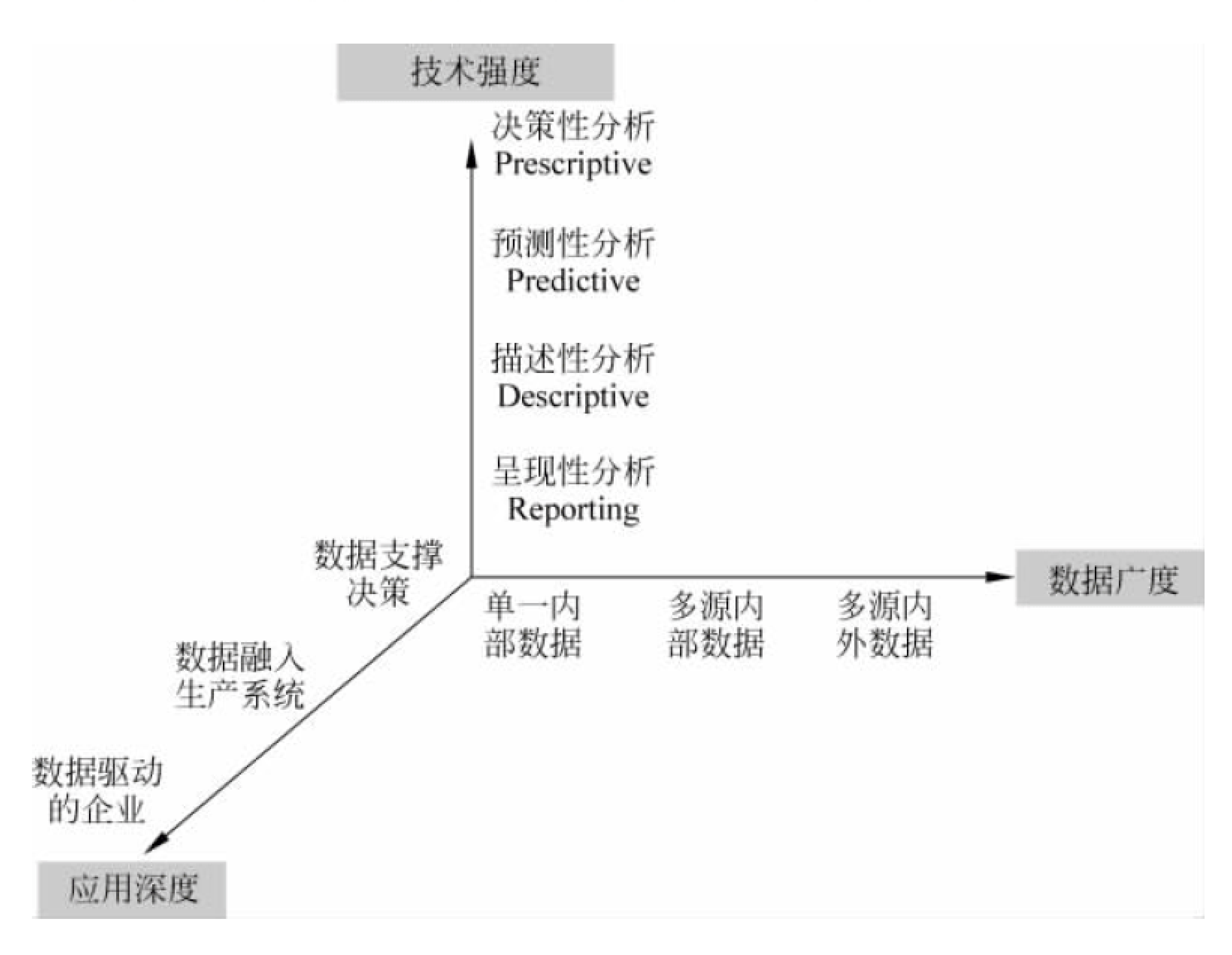


图 1-6 大数据应用的演进趋势

1. 可视化技术让数据平民化

大数据是一个由多维度、不断更新充实的数据群组成的，大数据的应用是需要从无规律、无直接因果的数据群中，根据需要或目标获得研究成果。数据的“可视化”则是大数据的展示手段。

最近几年，“大数据”概念深入人心。民众看到的大数据更多的是以可视化的方式体现的。可视化极大地拉近了大数据和普通民众的距离，即使对 IT 技术不了解的普通民众和非专业技术的常规决策者也能够很好地理解大数据及其分析的效果和价值，使得大数据可以从国计和民生两方面充分发挥其价值。

可视化是通过把复杂的数据转化为可以交互的图形，帮助用户更好地理解分析数据对象，发现、洞察内在规律。数据是人类对客观事物的抽象。人类对数据的理解和掌握是需要经过学习训练才能达到的。理解更为复杂的数据，必须越过更高的认知壁垒，才能对客观数据对象建立相应的心理图像，完成认知理解过程。好的可视化能够极大地降低认知壁垒，使复杂未知数据的交互探索变得可行。

可视化技术的进步和广泛应用对于大数据走向平民化的意义是双向的。一方面，可视化作为人和数据之间的界面，结合其他数据分析处理技术，为广大使用者提供了强大的理解、分析数据的能力。可视化使得大数据能够为更多人理解、使用，使得大数据的使用者从

少数专家扩展到更广泛的民众。另一方面,可视化也为民众提供了方便的工具,可以主动分析处理和个人工作、生活、环境有关的数据。民众服务的可视化技术,也将进一步和个人使用的移动通信设备(如手机)相结合。

2. 大数据安全与隐私令人忧虑

数据的增量呈指数增长,相应的大数据的安全问题也十分严峻。当大数据技术、系统和应用聚集了大量有价值的信息的时候,必将成为被攻击的目标。

大数据的过度滥用所带来的问题和副作用,最典型的就是个人隐私泄露。在传统采集分析模式下,很多隐私在大数据分析能力下变成了“裸奔”。类似的问题还包括,大数据分析能力带来的商业秘密泄露和国家机密泄露。

心理和意识上的安全问题,包括两个极端:一是忽视安全问题的盲目乐观,另一个是过度担忧所带来的对大数据应用发展的掣肘。比如,大数据分析对隐私保护的副作用,促使人们必须对隐私保护的接受程度有一个新的认识和调整。

大数据受到的威胁、大数据的过度滥用所带来的副作用、对大数据的极端心理,都会阻碍和破坏大数据的发展。

3. 新热点融入大数据多样化处理模式

大数据的处理模式依然多样化。大数据处理模式不断丰富,新旧手段不断融合,比如,流数据、内存计算成为新热点。内存计算继续成为提高大数据处理性能的主要手段。以 Spark 为代表的内存计算逐步走向商用,并与 Hadoop 融合共存。与传统的硬盘处理方式相比,内存计算技术在性能上有了数量级的提升。批处理计算、流计算、交互查询计算、图计算等多种计算框架使数据使用效率大大提高。很多新的技术热点持续地融入大数据的多样化模式中,目前还没有一个统一的模式。

4. 大数据提升社会治理和民生领域应用

基于大数据的社会治理成为业界关注的热点,涉及智慧城市、应急、税收、反恐、农业等多个民生领域。在最易获得大数据应用成果的互联网环境之后,大数据走进国计民生成为必然。

5. 深度分析推动大数据智能应用

在学术技术方面,深度分析会继续推动整个大数据智能的应用。这里谈到的智能强调涉及人的相关能力的延伸,比如决策预测、精准推介等,涉及人的思维和反射的延展、人的能力(智能和本能)的延展,这些都会成为大数据分析、机器学习、深度学习等学术技术发展的方向。

6. 数据权属与数据主权备受关注

数据成为重要的战略资源。人口红利、地大物博、经济实力、文化优势等都纷纷体现为数据资源储备和数据服务影响力。数据资源化、价值化是数据权属问题和数据主权问题的根源。

数据权属与数据主权被高度关注。大数据问题从个人和一般机构层面来看是数据权属问题,从国家层面来看是数据主权问题。数据的权属问题并不是传统的财产权、知识产权等可以涵盖的。数据成为国家间争夺的资源,数据主权成为网络空间主权的重要形态。

附：大数据年代记

1980 年，未来学家阿尔文·托夫勒将大数据称作“第三次浪潮的华彩乐章”。

2005 年，Hadoop 项目诞生，从技术层面上搭建了一个使对结构化和复杂数据快速、可靠分析变为现实的平台。

2008 年年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟(Computing Community Consortium)，发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不仅局限于数据处理的机器，并提出：大数据真正重要的是新用途和新见解，而非数据本身。此组织可以说是最早提出大数据概念的机构。

2009 年年中，美国政府通过启动 Data.gov 网站的方式进一步打开了数据的大门，这个网站向公众提供各种各样的政府数据。该网站超过 4.45 万的数据集被用于保证一些网站和智能手机应用程序跟踪从航班到产品召回再到特定区域内失业率的信息，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

2009 年，欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系，致力于改善在互联网上获取科学数据的简易性。

2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了专题报告《数据，无所不在的数据》。报告中提到：“世界上有着无法想象的巨量数字信息，并以极快的速度增长。从经济界到科学界，从政府部门到艺术领域，很多方面都已经感受到了这种巨量信息的影响。”库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2011 年 2 月，IBM 的沃森超级计算机每秒可扫描并分析 4TB(约两亿页文字量)的数据量，并在美国著名智力竞赛电视节目《危险边缘》(*Jeopardy*)上击败两名人类选手而夺冠。《纽约时报》认为这一刻为一个“大数据计算的胜利。”

2011 年 5 月，全球知名咨询公司麦肯锡(McKinsey&Company)全球研究院(MGI)发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，从此大数据开始备受关注，这也是专业机构第一次全方面地介绍和展望大数据。报告指出，大数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。报告还提到，“大数据”源于数据生产和收集的能力和速度的大幅提升——由于越来越多的人、设备和传感器通过数字网络连接起来，产生、传送、分享和访问数据的能力也得到彻底变革。

2011 年 12 月，中国工业和信息化部发布的物联网“十二五”规划上，把信息处理技术作为 4 项关键技术创新工程之一被提出来，其中包括海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。

2012 年 1 月，瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响》(*Big Data, Big Impact*)宣称，数据已经成为一种新的经济资源类别，就像货币或黄金一样。

2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。奥巴马政府宣布两亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭。奥巴马政府将数据定义为“未来的新

石油”，大数据技术领域的竞争，事关国家安全和未来。国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用的能力；国家数字主权体现对数据的占有和控制。数字主权将是继边防、海防、空防之后，另一个大国博弈的空间。

2012 年 4 月 19 日，美国软件公司 Splunk 在纳斯达克成功上市，成为第一家上市的大数据处理公司。Splunk 成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快大数据布局。

2012 年 7 月，联合国在纽约发布了一份关于大数据政务的白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求：例如，通过对价格关注和更好服务的渴望，个人提供数据和众包信息，并对隐私和退出权力提出需求；公共部门出于改善服务、提升效益的目的，提供了诸如统计数据、设备信息、健康指标，及税务和消费信息等，并对隐私和退出权力提出需求；私人部门出于提升客户认知和预测趋势的目的，提供汇总数据、消费和使用信息，并对敏感数据所有权和商业模式更加关注。白皮书还指出，人们如今可以使用的极大丰富的数据资源，包括旧数据和新数据，来对社会人口进行前所未有的实时分析。如果政府能合理分析所掌握的数据资源，将能“与数俱进”，快速应变。

2012 年 7 月，为挖掘大数据的价值，阿里巴巴集团在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台——“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。马云强调：“假如我们有一个数据预报台，就像为企业装上了一个 GPS 和雷达，你们出海将会更有把握。”因此，阿里巴巴集团希望通过分享和挖掘海量数据，为国家和中小企业提供价值。此举是国内企业最早把大数据提升到企业管理层高度的一次重大里程碑。阿里巴巴也是最早提出通过数据进行企业数据化运营的企业。

2012 年 12 月，英国数据战略委员会成立了世界上第一个非盈利性的开放数据协会(Open Data Institute，ODI)，推动开放数据的进程。

2013 年，“开放政府联盟(OGP)”的 8 个成员国(美国、英国、法国、德国、意大利、加拿大、日本及俄罗斯)签署《开放数据宪章》，承诺在 2013 年年底前，制定开放数据行动方案。截止到 2014 年 2 月，全球已有 63 个国家加入 OGP。

2014 年 4 月，世界经济论坛以“大数据的回报与风险”主题发布了《全球信息技术报告(第 13 版)》。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要。在接下来将对数据保密和网络管制等议题展开积极讨论。全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到大数据在推动经济发展、改善公共服务、增进人民福祉，乃至保障国家安全方面的重大意义。

2014 年 5 月，美国白宫发布了 2014 年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。

2014 年 8 月，联合国开发计划署首次携手科技企业共建大数据实验室，利用大数据技术和联合国的全球发展经验，在环境保护、医疗与疾病预防、教育、扶贫等诸多领域进行深入

的研究分析，推动大数据解决全球问题的创新模式，促进持续发展。

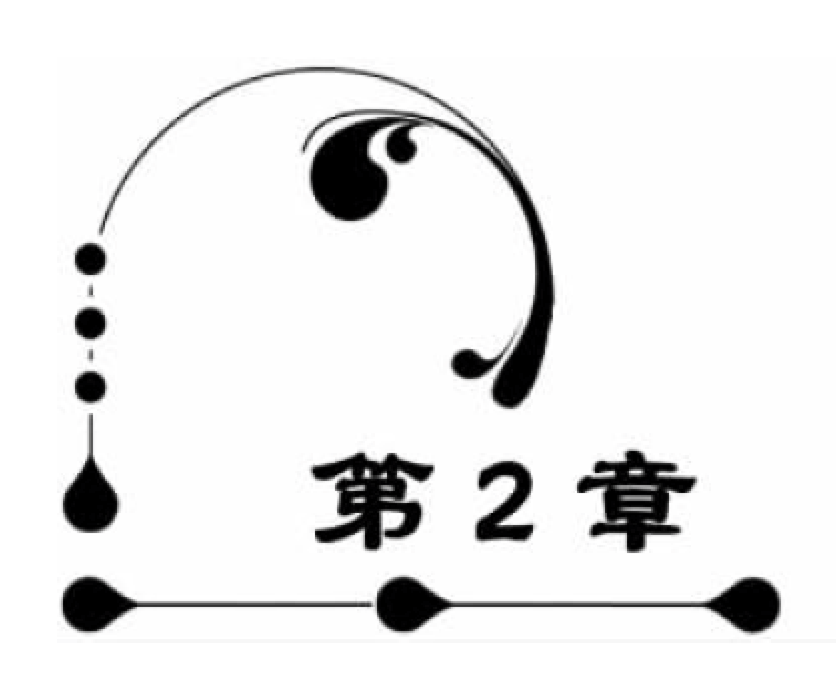
2014 年，世界经济论坛以“大数据的回报与风险”为主题发布了《全球信息技术报告(第 13 版)》。

2015 年 8 月 31 日，中华人民共和国国务院印发《促进大数据发展行动纲要》(国发〔2015〕50 号)。

2016 年 8 月 26 日，《大数据产业“十三五”发展规划》已经呼之欲出。此次编制的《大数据产业“十三五”发展规划》中，工业大数据、大数据资源开放共享、大数据交易、大数据安全、大数据标准、大数据行业应用等领域是研究重点。

2016 年 11 月 17 日，全球大数据高峰论坛在中国青岛隆重召开。以“数据创造价值·智慧引领未来”为主题，聚焦全球大数据发展现状，致力于发现大数据与产业融合之道，挖掘大数据行业发展的未来机遇，推动大数据应用研究与信息经济进一步发展壮大。

2017 年 1 月 17 日，中国工业和信息化部印发了《大数据产业发展规划(2016—2020 年)》提出，到 2020 年，大数据相关产品和服务业务收入突破 1 万亿元，年均复合增长率保持在 30%左右。



第 2 章 大数据关键技术

大数据的概念和特征都说明了大数据不是一个简单的名词，而是一系列技术综合运用的集合。大数据处理关键技术一般包括大数据采集、大数据预处理、大数据存储及管理、大数据分析及挖掘、大数据展现和应用(大数据检索、大数据可视化、大数据应用、大数据安全等)。

大数据产生有其必然性，主要归结于互联网、移动设备、物联网和云计算等的快速崛起，全球数据量大幅提升，是实现大数据采集、存储、处理和呈现的有力武器，在很大程度上是大数据产生的原因，应用需求引导着技术的研发方向，技术突破促进了创新模式的实现。

如果把新一代信息系统类比成人体：物联网是“感官”；移动互联网是“神经”；云计算是“心脏和体魄”；大数据是“聪明的大脑”，如图 2-1 所示。

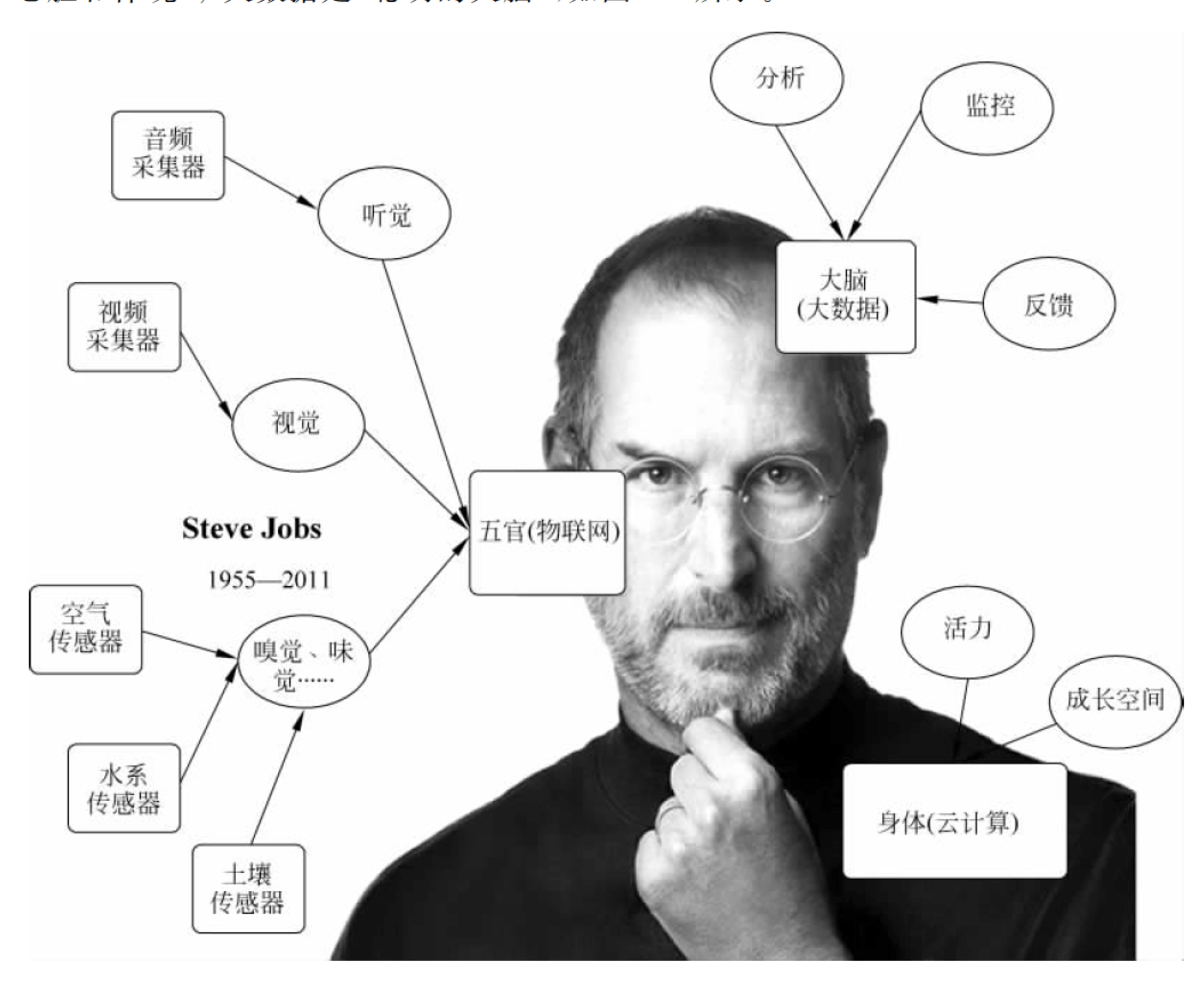


图 2-1　大数据“聪明的大脑”

2.1 物联网

人类从古到今，一直面临一个巨大的问题：如何获取信息、传递信息、处理信息和应用信息。面对周围复杂的环境，人们需要各种各样的信息方便工作和学习，以便提高效率。从古代的驿站飞鸽传书到今天的发报机、固定电话、手机、对讲机、卫星电话以及互联网都是为了解决信息的传递。随着微电子技术和自动化控制技术的发展、电子计算机的产生，人类处理信息的能力大大提升。

但是，随着社会的发展，人们需要获取的信息飞速增长，如何更有效地获取信息、传递信息和使用信息成为迫切要解决的问题。物联网技术的出现为人们解决数据采集问题绘制了一幅宏大的蓝图。

2.1.1 物联网的概念

1. 物联网概念提出（1995—1999 年）

物联网（Internet of Things，IoT）概念起源于比尔·盖茨 1995 年所编写的《未来之路》一书。在该书中，比尔·盖茨已经提及物联网的概念，只是当时受限于无线网络、硬件及传感设备的发展，并未引起人们的重视。

1998 年，美国麻省理工学院（MIT）创造性地提出了当时被称作 EPC（Electronic Product Code）系统的“物联网”的构想。1999 年，美国麻省理工学院 Auto_ID 中心首先提出“物联网”的概念，即将所有物品通过射频识别等信息传感设备与互联网连接起来，实现智能化识别和管理的网络。人们将按照特定的数据格式，将每一件物品赋予一个唯一的编号，这个编号就是 EPC，而电子标签是这一编号的载体。基于互联网和射频技术的 EPC 系统，即实物物联网（简称物联网）是在计算机互联网的基础上，利用 RFID（Radio Frequency Identification，射频识别）、无线数据通信等技术构造了一个实现全球物品信息实时共享的网络。

该阶段物联网的概念内涵很小，具体等同于物品的联网，目的是实现全球物品的信息实时共享。其主要组成部分包括 RFID、无线数据通信、互联网和数据存储。

2. 概念延伸（1999—2008 年）

2005 年 11 月 17 日信息社会世界峰会（World Summit on the Information Society，WSIS）上，国际电信联盟发布了《ITU 互联网报告 2005：物联网》，正式提出了“物联网”的概念（如图 2-2 所示），包括人与物、物与物之间的连接，即在任何时间、任何地点、任何物品间都可以进行通信，重点突出了连接对象的无所不包和网络的无处不在以及物体的智能化，重点体现把“物”纳入网络中和任何地方都有网络。

ITU 的报告描绘了“物联网”时代的图景：当司机出现操作失误时汽车会自动报警；公文包会提醒主人忘带了什么东西；衣服会“告诉”洗衣机对颜色和水温的要求等。

2008 年 9 月 5 日，EPOSS（European Technology Platform on Smart Systems Integration，欧洲智能集成系统技术平台）在《2020 年的物联网：未来发展方向》中指出，物联网将通过智能接口连通社会、环境和用户，形成一个智慧的空间，其中的任何事物都具有一个独立的虚拟的身份标识，并且任何人在其中都可以进行个性化的操作。

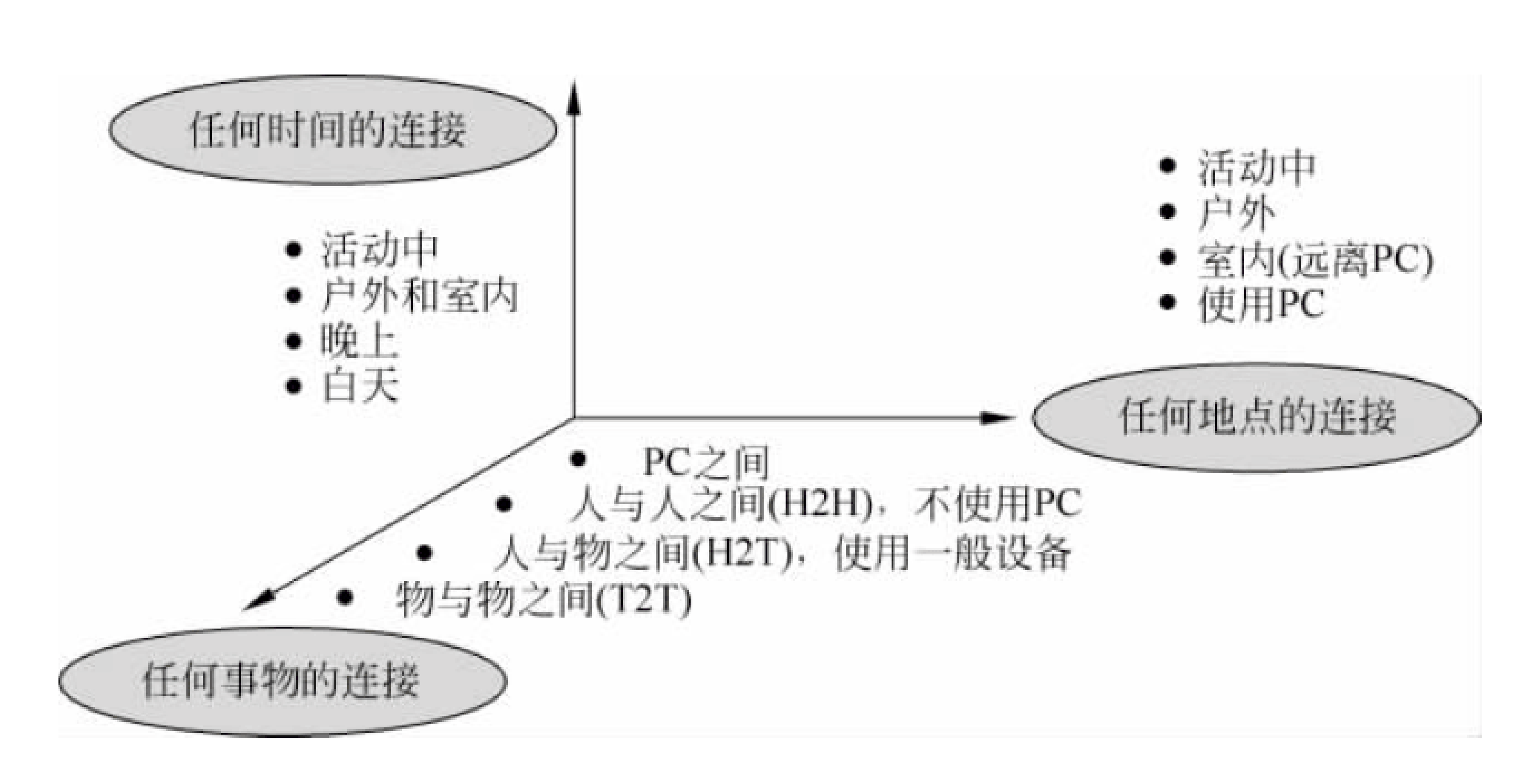


图 2-2　ITU 对物联网的界定

（资料来源：《ITU 互联网报告 2005：物联网》）

2008 年 10 月，巴黎高等电信商学院的多位专家在 *The Internet of Things—What Challenges for Europe*？中给出了物联网的建议性定义：物联网是可以方便识别数字实体和物理对象的相互连通的网络，无论是无生命的（基础设施）或有生命的（动物、植物和人）都包含在内，通过标准的电子识别系统和无线通信设备，都可以获取、存储、传输和处理各类信息数据，将现实与虚拟无缝连接在一起。

欧盟委员会则认为物联网将会是未来网络的整合，它是以标准的、互通的通信协议为基础，具有自我配置能力的全球性动态网络。在这个网络中，所有实体的和虚拟的物体都有特定的编码和物理特性，通过智能界面无缝连接，实现信息的共享。

物联网的概念得到了进一步延伸，从物联网的对象上来说，人既是物联网中信息的使用者和受益者，也是物联网的连接对象，也就是说，在物联网中不仅包括"物"，还包括"人"，人已经被看成了"物品"。从应用的范围上来说，物联网可以实现任何时间、任何地点以及任何人和物的通信。同时将射频识别技术（RFID）、传感器技术、纳米技术、智能嵌入技术指定为物联网发展的 4 大关键技术。

3. 物联网概念的界定

随着 IBM 提出"智慧地球"和美国政府的积极响应，全球很多国家将物联网上升到国家战略角度，并制定了相关的信息战略。如美国的"智慧地球"；欧盟的"超越 RFID——物联网"；日本的"泛在信息社会"；韩国的"物联网规划"；新加坡的"智慧国 2015"计划。我国国务院出台的《关于推进物联网有序健康发展的指导意见》进一步明确发展目标和发展思路，推出 10 个物联网发展专项行动计划落实具体任务。越来越多的各国学者、企业家、研究机构和政府机构也开始或加大对物联网概念、组成、应用以及创造的产业价值等各方面的研究。物联网内涵的界定如图 2-3 所示。

狭义的物联网就是物和物的通信，也就是图 2-3 中右边虚线圈内的部分，如果我们把人也看做物，那么整个图就是广义的物联网。

物联网是在计算机互联网的基础上，利用 RFID 等技术构造一个覆盖世界上万事万物的"Internet of Things"。通过计算机互联网实现物品（商品）的自动识别和信息的互联与共享。

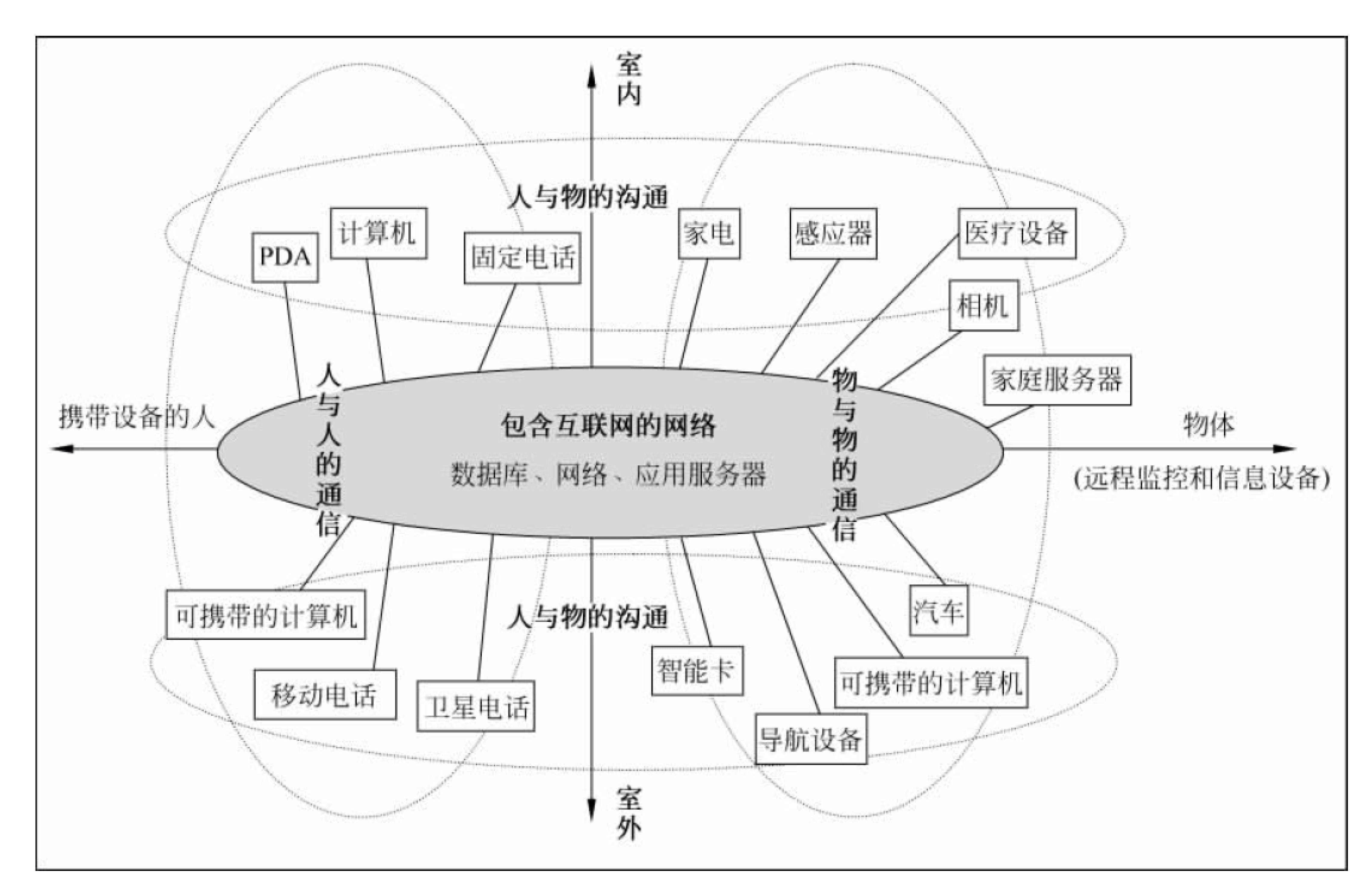


图 2-3 物联网的含义

2.1.2 物联网：大数据资源的重要提供者

物联网的应用是基于物联网自身的特点而发展起来的。物联网具有以下特点。

（1）对物品实现唯一标识；

（2）对物品快速分级进行处理；

（3）对物品物流信息实行实时监控；

（4）对信息进行非接触自动处理；

（5）可以实现各环节信息共享。

基于这些特点，物联网在许多行业发展了不同的应用，比如物流业方面的供应链管理、安全检测方面的环境监测、医疗业方面的电子病历、灾害/危机管理等，如图 2-4 所示。

以物流业为例，在供应链管理中每个商品都有唯一的编码对其进行编号，然后根据这个编号实时监控商品位置，以确定商品的传输流程。由于商品一直在移动、半移动状态，所以只能对商品进行非接触式自动处理，同时由于商品信息需要在各个部门之间流动，因此需要信息的共享以便能进行有效的管理。

工业方面，以美国工业互联网、德国工业 4.0 和中国制造 2025 战略为代表，物联网成为实现制造业智能化变革和重塑国家竞争优势的关键技术基础，围绕物联网的全球生态构建和产业布局正加速展开。

同时，物联网也是智慧城市发展的核心基础要素，在城市管理、节能减排、能源管理、智能交通等领域进行广泛应用，“前端设备智能化＋后端服务平台化＋大数据分析”成为通用模式[①]。

① 2015 物联网白皮书[R]. 中国信息通信研究院，2015.

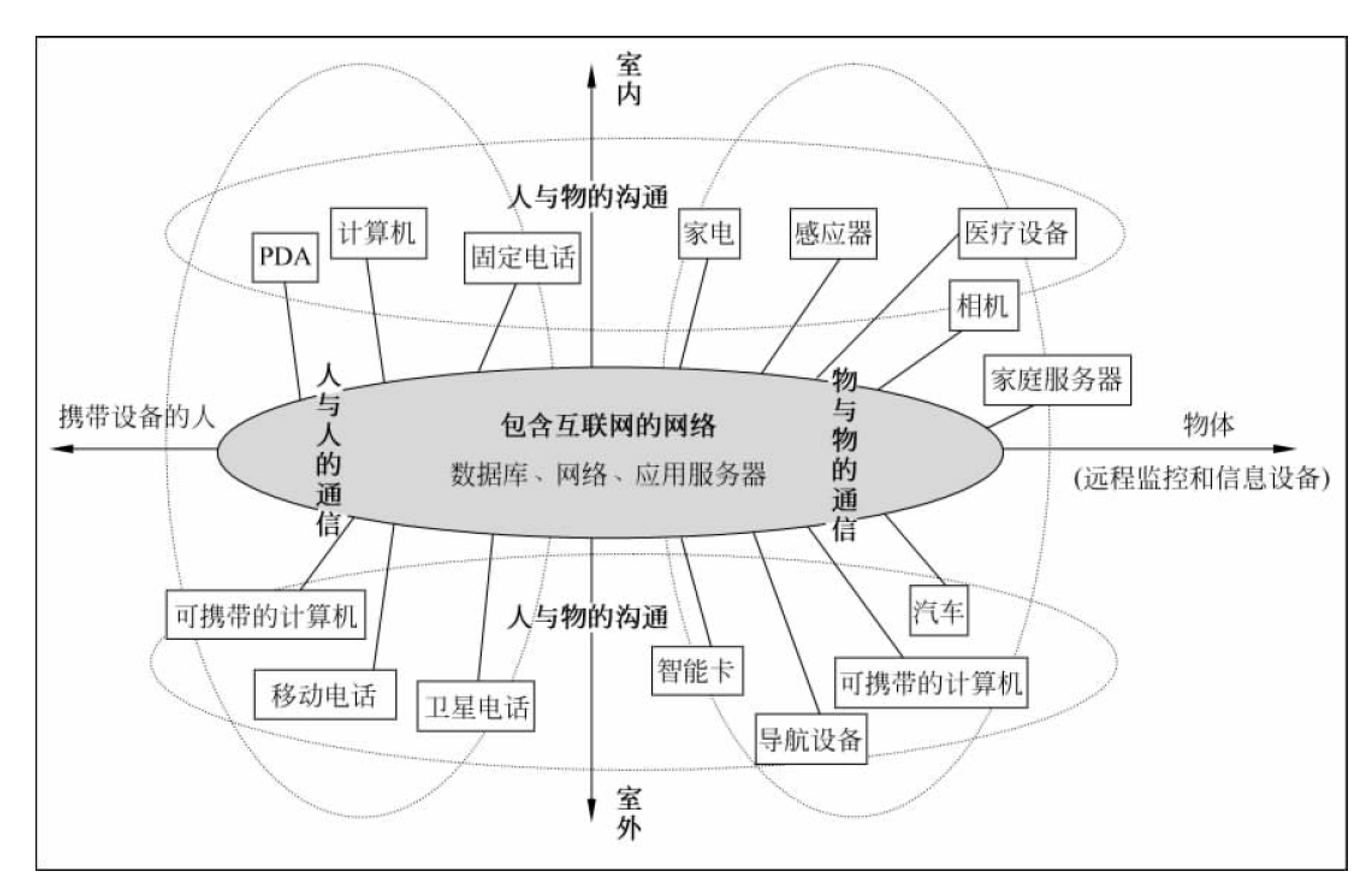


图 2-3 物联网的含义

2.1.2 物联网：大数据资源的重要提供者

物联网的应用是基于物联网自身的特点而发展起来的。物联网具有以下特点。

（1）对物品实现唯一标识；

（2）对物品快速分级进行处理；

（3）对物品物流信息实行实时监控；

（4）对信息进行非接触自动处理；

（5）可以实现各环节信息共享。

基于这些特点，物联网在许多行业发展了不同的应用，比如物流业方面的供应链管理、安全检测方面的环境监测、医疗业方面的电子病历、灾害/危机管理等，如图 2-4 所示。

以物流业为例，在供应链管理中每个商品都有唯一的编码对其进行编号，然后根据这个编号实时监控商品位置，以确定商品的传输流程。由于商品一直在移动、半移动状态，所以只能对商品进行非接触式自动处理，同时由于商品信息需要在各个部门之间流动，因此需要信息的共享以便能进行有效的管理。

工业方面，以美国工业互联网、德国工业 4.0 和中国制造 2025 战略为代表，物联网成为实现制造业智能化变革和重塑国家竞争优势的关键技术基础，围绕物联网的全球生态构建和产业布局正加速展开。

同时，物联网也是智慧城市发展的核心基础要素，在城市管理、节能减排、能源管理、智能交通等领域进行广泛应用，“前端设备智能化＋后端服务平台化＋大数据分析”成为通用模式[①]。

---

① 2015 物联网白皮书[R]. 中国信息通信研究院，2015.



图 2-4　物联网的应用

物联网的快速发展，使其成为大数据资源重要的提供者。相对于现有互联网数据的杂乱无章和价值密度低的特点，通过可穿戴、车联网等多种数据采集终端定向采集的数据资源更具利用价值。

智能化的可穿戴设备层出不穷，例如智能手环、腕带、手表等可穿戴产品正在走向成熟，智能钥匙扣、自行车、筷子等，国外的 Intel、Google、Facebook，国内的百度、京东、小米等企业都在该领域内有所布局。根据 IDC 公司统计，到 2016 年年底，全球可穿戴设备的出货量达到一亿多台，较 2015 年增长 29.0%。到 2020 年之前，可穿戴设备市场的年复合增长率将为20.3%，而 2020 年将达到 2.136 亿台[①]。可穿戴设备可以实现 7×24 小时不间断地收集个人健康数据，在医疗保健领域有广阔的应用前景，一旦技术成熟，设备测量精度达到医用要求，电池续航能力也有显著增强，就很可能进入大规模应用阶段，从而成为重要的大数据来源。

此外，据 Strategy Analytics 公司预计，车载前端联网模式在未来 5 年内将迎来发展黄金期，2020 年将达到 49%[②]。车联网的成熟发展，将为基于位置及路况的大数据应用提供更为实时、准确的基础数据资源。

2.2　移动互联网

伴随着互联网技术的完善和移动通信技术的不断升级换代，移动互联网作为移动通信和传统互联网融合的产物，被视为未来网络发展最重要的趋势之一。移动互联网的发展颠覆了世界的方方面面。

2.2.1　移动互联网的发展

2015 年 3 月世界通信大会发布《移动经济 2015》，报告指出：大量移动通信用户开始享用 3G 及 4G 宽带网络，移动宽带（3G＋4G）通信用户比例已达 40%，预计到 2020 年将增至

① http://www.idc.com/getdoc.jsp? containerld=prUS41530816.

② http://www.askci.com/new/dxf/20160727/15510447326.shtml.

约70%①。2015年12月1日，国际电信联盟(ITU)发布了年度互联网调查报告。报告显示，全球手机用户数达到71亿，手机信号已覆盖了全球超过95%的人口，已有32亿人联网。

中国互联网络信息中心(CNNIC)2017年1月发布的第39次《中国互联网络发展状况统计报告》显示，截至2016年12月，中国网民规模达7.31亿，普及率达53.2%，超过全球平均水平3.1个百分点，超过亚洲平均水平7.6个百分点。全年共计新增网民4299万人，增长率为6.2%。其中，手机网民规模达6.95亿，占比达95.1%，增长率连续三年超过10%②，如图2-5所示。

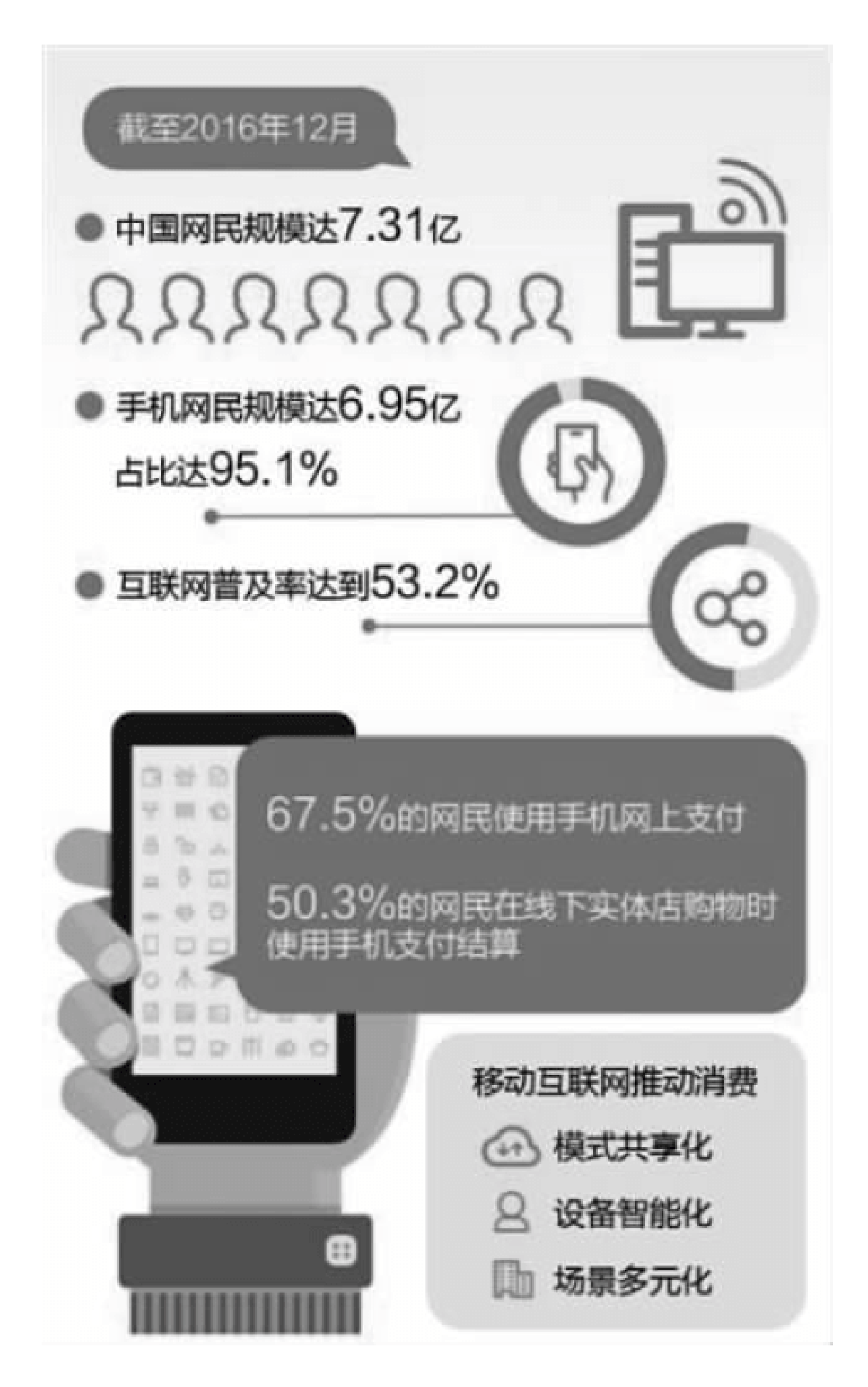


图2-5 移动互联网

移动互联网已经形成一个超过万亿美元规模的巨大产业，并迅速应用于金融、商务、物流、医疗、教育等社会各行业，对经济社会的影响急速放大，乃至成为“互联网+”的基础设施。移动互联网时代已经到来！

移动互联网体现了“无处不在的网络、无所不能的业务”的思想，正在改变着人们的生活方式和工作方式。移动互联网应用具有移动性和个性化等特征：用户可以随时随地获得移动互联网服务；这些服务可以根据用户位置、兴趣偏好、需求和环境进行定制。随着大数据、云计算等技术的发展，用户从信息的获得者变为信息的贡献者，基于群体用户、个人用户需求、位置等信息的深度挖掘，移动互联网应用趋于个性化和智能化。

移动互联网具有带动其他产业发展和附加值高以及运营总体成本很低的优势，也极大

① 2015移动互联网白皮书[R]. 中国信息通信研究院，2015.

② 中国互联网络信息中心(CNNIC). 中国互联网络发展状况统计报告(第39次)[R]. 2017.

地促进了企业能够快速便捷地发展自身的电子商务等其他方面的产业。这种发展模式已经开始潜移默化地渗透到很多行业,加快了传统行业向全新的经营模式转变的脚步[①]。

2.2.2　移动互联网:大数据的传输载体

移动互联网补充了传统互联网的网络空隙,形成了一个全覆盖的泛在网络空间,是承载和传送各类数据的“邮差”。大数据的信息收集与传输需要无线电技术作为其载体,如图 2-6 所示。

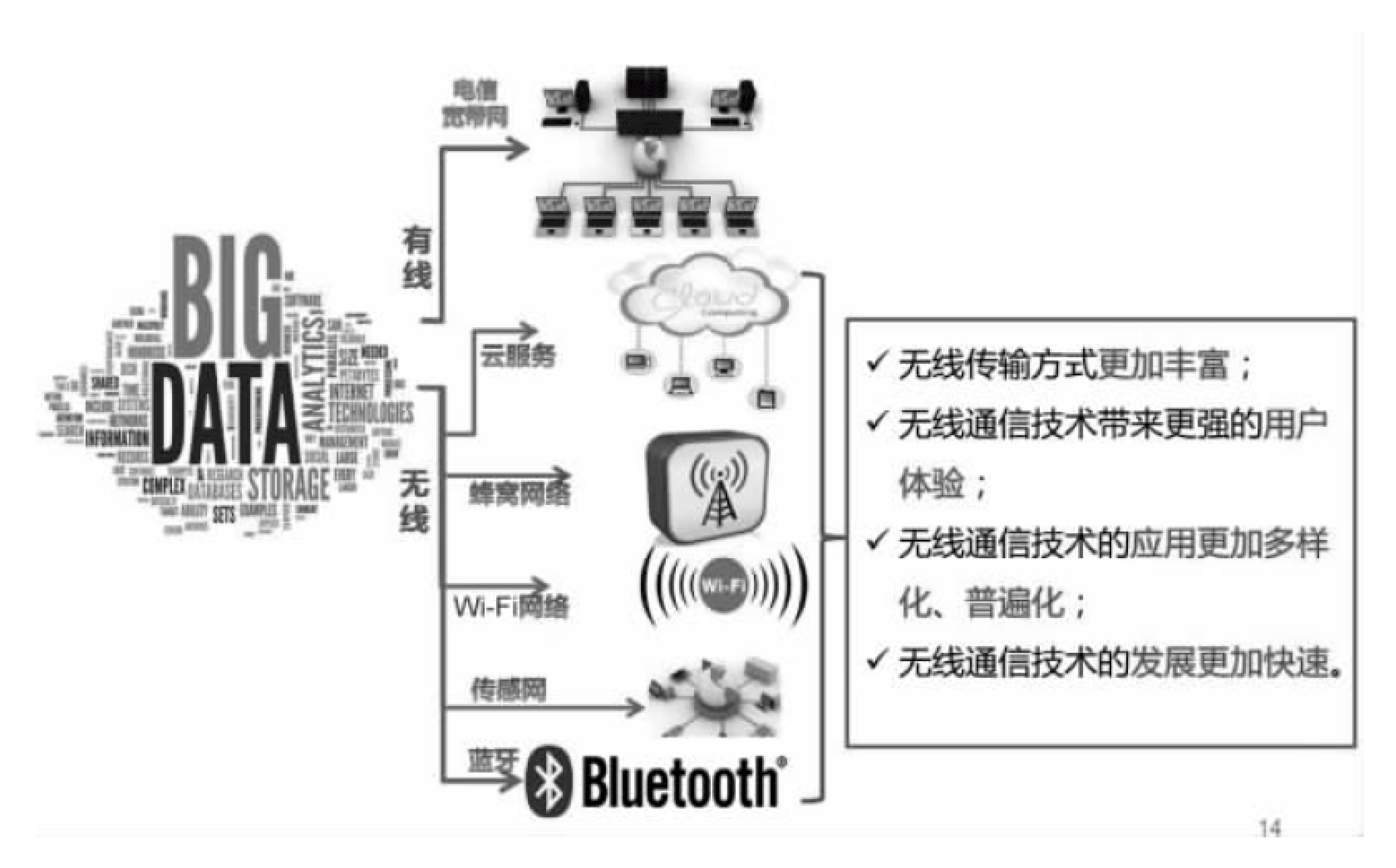


图 2-6　大数据的传输载体

大数据的传输载体按照技术方式的不同可以大致分为两类:有线传输接入,无线传输接入。其中,有线传输接入以宽带为代表,而无线传输接入则有广电网、无线电信网、WAPI/Wi-Fi、蓝牙、WLAN、WiMax、ZigBee 等方式。各类无线技术的传输特性如图 2-7 所示。

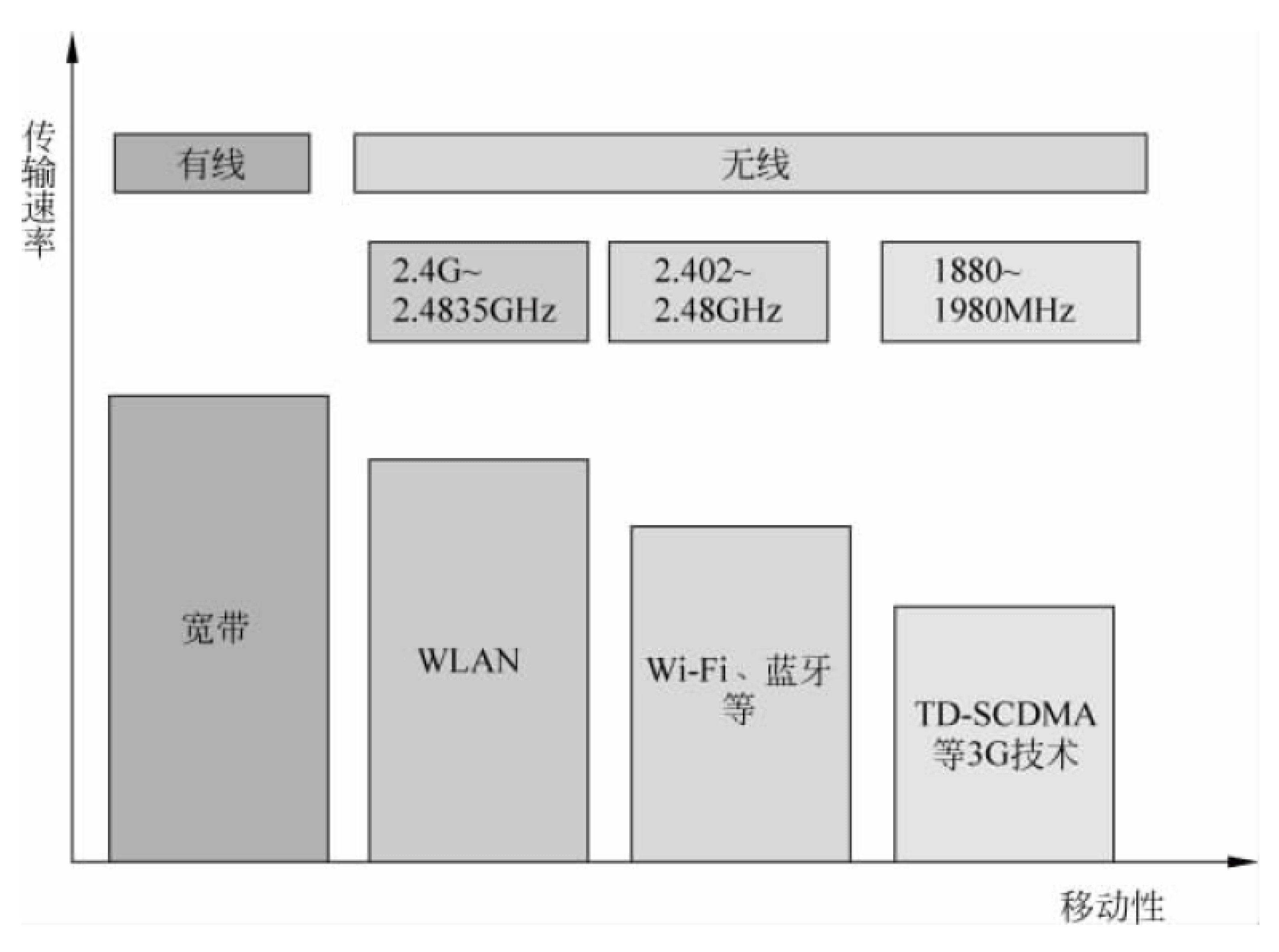


图 2-7　常用无线技术的传输特性

① 李春生. 移动互联网发展趋势研究[J]. 中国高新技术企业,2016,(01):1-2.

不同传输技术之间存在着传输速率与移动性的差异，而传输速率和移动性又相互形成互补关系，使得各项技术能够在物理空间实现无缝连接，一些移动性不好的技术能够有较高的传输速率，而移动性好的技术移动速率则偏低，最终使得在任何时间、任何空间都能实现物联网的传输与接入。

2.3 云计算

DT(数据技术)时代将是客户体验至上的时代，这需要数据计算能力的提升，大数据、数据挖掘、数据分析等工具为发现和利用这些数据的价值带来了可能，数据正在被云计算发展和征服。

云计算是并行计算(Parallel Computing)、分布式计算(Distributed Computing)和网格计算(Grid Computing)的发展，或者说是这些计算机科学概念的商业实现。云计算是虚拟化(Virtualization)、效用计算(Utility Computing)、IaaS(基础设施即服务)、PaaS(平台即服务)、SaaS(软件即服务)等概念混合演进并跃升的结果。

2.3.1 云计算的优点

云计算使计算分布在大量的分布式计算机上，而非本地计算机或远程服务器中，用户数据中心的运行与互联网更相似。这使得用户能够将资源切换到需要的应用上，根据需求访问计算机和存储系统，数据通过互联网或相应专属网络进行传输。以云计算为基础的信息处理、存储和分享手段，可以更便捷、高效地将这些大量、高速、多种数据类型、价值密度低的数据进行存储、传输与分析计算。应用云计算技术对大数据进行计算、分析，可以发掘和释放出更多的数据隐藏价值，为实现精准决策提供更多的有用信息。

云计算具有如下优点。

(1) 超大规模，特高性能。现有提供云服务的企业如Google、Amazon、IBM、微软、Yahoo、阿里云等，都拥有超过数十万台的云计算服务器。一般企业私有云也具有成百上千台服务器。云计算服务器在大规模布置的同时也为用户赋予了前所未有的计算性能，原来需要耗费大量时间、大量资源的计算任务，通过使用云计算服务可以极大地降低计算时间和所耗资源。

(2) 虚拟化，高弹性。云计算将传统的计算、网络和存储资源，通过使用虚拟化、容错和并行处理等方式，转化为可以弹性伸缩的服务，用户可以在任意位置、通过各种终端获取应用服务而无须考虑应用运行的具体位置和运行方式。

(3) 服务可靠性高。云计算通过多副本容错、计算节点同构可互换等措施来保障服务的高可靠性。添加、删除、修改云计算环境的任一资源节点，抑或任一资源节点异常宕机，都不会导致云计算环境中的各类业务的中断，也不会导致用户数据的丢失。这里的资源节点可以是计算节点、存储节点和网络节点。

(4) 资源动态扩展。云计算整合众多计算机资源，构成技术存储模式，实现并行计算、网格计算、分布式计算、分布式存储等多种方式。在云计算平台下通过资源调度机制动态控制云计算规模，满足应用和计算规模增长的需要。在系统业务需求升高时，启动闲置资源纳入系统，提高云平台的承载能力；在系统业务负载低时，将业务集中起来，释放部分资源闲置给云之外的其他应用：从整体上动态调整资源使用情况。

(5) 按使用付费,价格廉价。云计算平台通过虚拟分拆技术实现了计算资源的同构化和可度量化。云计算服务提供一个庞大的资源池,用户可以按需购买,而且目前大多云服务采用的是按使用量计费的收费模式,用户仅需对所使用的资源付费。此外,“云”的可动态扩展资源的特点也使其资源利用率较传统系统有了大幅提升,而且“云”的自动化集中式管理也使大量企业无须负担高昂的数据中心管理成本,用户可以充分享受云服务带来的价廉而高性能的服务。

人类进入了数字时代,数字化带给人类的是史无前例的效率提升和资源节约。但是,信息化的历史告诉我们,我们在发展信息化方面已经造成的浪费可能比在其他方面浪费得更多。从大数据概念的提出到现在已经过去了七八年的时间,可是人们对于大数据、数据挖掘、数据统计和数据分析仍然没有分得很清楚,分不清在何种环境下的哪些问题需要采用什么样的技术去处理,分不清在什么情况下传统的技术已经无法解决问题而必须发展云计算。模糊的认识必然影响大数据的利用和发展,很可能形成为了节约却造成了更大浪费的情况。一方面是世界技术的高速发展,另一方面是我们较低的数据应用水平和能力,这是我国当前大数据研究和发展方面面临的最大问题。

2.3.2 云计算与大数据的关系

大数据与云计算的关系就像一枚硬币的正反面一样密不可分,一个是问题,一个是解决问题的必然方法。大数据必然无法用单台的计算机进行处理,必须采用分布式计算架构。大数据的特色在于,对海量数据的运算必须依托云计算的分布式处理、存储与访问技术。正是因为云计算采用分布式运算技术,云计算的速度和实时性较传统技术要高很多。云计算解决的是数据的处理能力和处理效率,与大数据技术相辅相成,如图 2-8 所示。



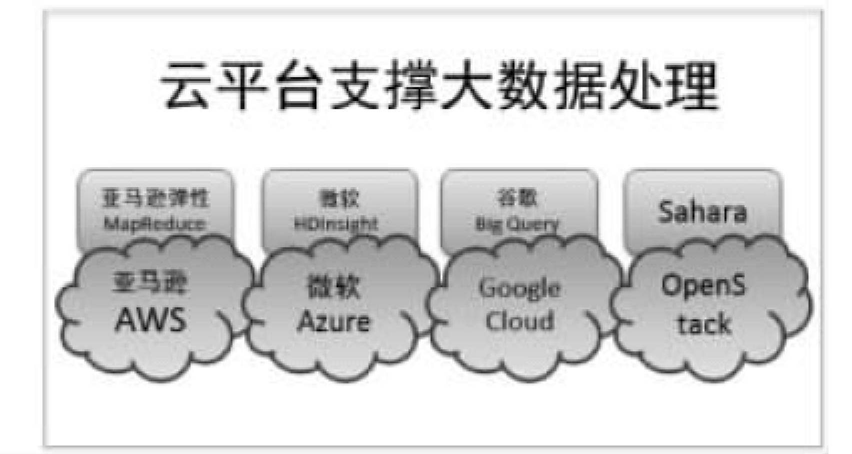


图 2-8 云计算与大数据

多样的数据类型涉及数字、文字、图片、语音、视频、地理位置、网络日志信息等。从数据结构来看,可分为非结构化数据、半结构化数据和结构化数据;从数据存储方案的角度,还可以分为分布式存储和集中式存储;从数据质量来看,数据的完整性、可信性与可用性也大不相同。对于是否必须采用大数据运算,在数据规模和数据复杂性之间存在一定的取舍关系。一般来说,分布式存储就意味着很大的数据体量,分布存储的数据就需要用大数据技术

来处理了，传统技术已经无法使用。大数据适合于处理分布式存储的复杂数据。

新型的处理技术主要是指云计算技术。最典型的云计算技术是以 Google 公司的 Hadoop 为代表的云计算技术，其中包括 HDFS、Hadoop Map Reduce 以及 Hadoop Common。

云计算技术出现之前，传统的计算机、数据库完全无法处理如此量大且不规则的非结构数据。云计算技术的出现使大数据的处理成为可能，使得原来在有限的服务器、数据库、终端上需要耗费大量的人力、物力和时间才能完成的运算问题不再艰难。可以说大数据运算就意味着：运用云计算技术，高速处理体量在 10TB 级别以上的分布式数据或各类复杂数据。

2.4 智慧旅游的大数据采集

近些年，随着假日经济的发展，我国旅游业发展进入了上升通道，在繁荣发展的背后问题也逐渐显现出来。游客的数量迅速增长也给景区、交通等各个相关行业带来压力。在热门旅游时间点，如十一等旅游黄金周，多次出现了部分景点人满为患的现象，提高了景区管理的难度，也影响了游客的体验感知。

大数据不同于传统的数据应用，主要体现在数据的“体量巨大、多维度、实时性”，其中，“多维”一方面指面向同一主体的数据视角丰富，另一方面也指数据结构上不同于传统的结构化数据，是包括图像、视频等数据在内的多类型数据的整合应用。

旅游景区的大数据应用，相对互联网企业，其数据量相对较小，其应用的难点更多在于采集处理“多维度”数据和“实时性”数据两个方面。

旅游景区大数据的多维度包括景区 IT 平台原有数据、视频监控数据、各类景区传感器数据、地理信息数据、气象数据、外部互联网数据等。其中，互联网数据又涵盖了 LBS 用户定位、搜索数据、网上交易数据、社交数据等，这些数据都可以通过科学的数据模型进行整合梳理找出有价值的规律和方向。旅游景区大数据的实时性主要体现在基于视频、GPS 等捕捉到的监控数据，用于客流监控、轨迹监控和防火预警等方面。

数据源汇聚包括景区内部采集沉淀的多类数据以及外部引入的数据两类数据资源整合，以及搭建统一的信息平台承载相应的数据资源两个重要部分。

2.4.1 整合内外部数据

1. 物联网采集的内部数据

是多层次感知景区的重要技术手段，是通过射频识别（RFID）、红外感应器、全球定位系统、激光扫描器、二维码识别终端等信息传感设备，按约定的协议把各类物品和互联网连接起来，进行信息交换和通信，以实现智能化识别、定位、跟踪、监控和管理的一种网络。

物联网实现了人与人、人与机器、机器与机器的互联互通。通过 RFID、传感器、二维码等信息传感设备植入门票、桥梁、公路、建筑、供水系统、电网等景区的各种物体中，可以实现对景区更透彻的感知；通过与互联网的融合，能将景区事物信息实时准确地传递出去，从而实现更为广泛的互联互通；通过利用云计算、模糊识别等各种智能计算技术，对海量的数据和信息进行分析和处理，能够帮助对景区内各类人和物实施智能化的控制，如图 2-9 所示。



图 2-9　智慧景区物联网技术应用

2. 引入互联网数据等外部数据资源

互联网尤其是移动互联网的兴起，使得线上信息总量正以极快的速度不断暴涨。每天在微博、微信、论坛、新闻评论、电商平台上分享各种文本、照片、视频、音频、数据等信息高达几百亿甚至几千亿条，这些信息涵盖着商家信息、个人信息、行业资讯、产品使用体验、商品浏览记录、商品成交记录、产品价格动态等海量信息。这些数据通过聚类可以形成旅游行业大数据，其背后隐藏的是旅游行业的市场需求、竞争情报，闪现着巨大的财富价值。

目前，围绕开放数据源的产业生态逐步形成，有大量的公司从事互联网等开放数据的采集分析。同时，随着大数据应用对外部数据需求的增加，国内近些年也出现了包括数据堂、贵阳大数据交易所在内不同类型的数据交易机构，使得在大数据应用的过程中，能够较为高效地获得外部数据。本案例通过与外部公司合作方式获得了游客在各个平台生成的数据，比如旅游攻略相关网站、社交平台、位置数据等。

2.4.2　信息化平台——数据采集存储基础设施

如图 2-10 所示为旅游大数据信息化平台。

1. 信息基础实施

主要指各种传感设备（射频传感器、位置传感器、能耗传感器、速度传感器、热敏传感器、湿敏传感器、气敏传感器、生物传感器等），这些设备嵌入到景区的物体和各种设施中，并与互联网连接。

2. 统一的数据中心

数据中心是景区信息资源数据库的存储中心、管理服务中心和数据交换中心。

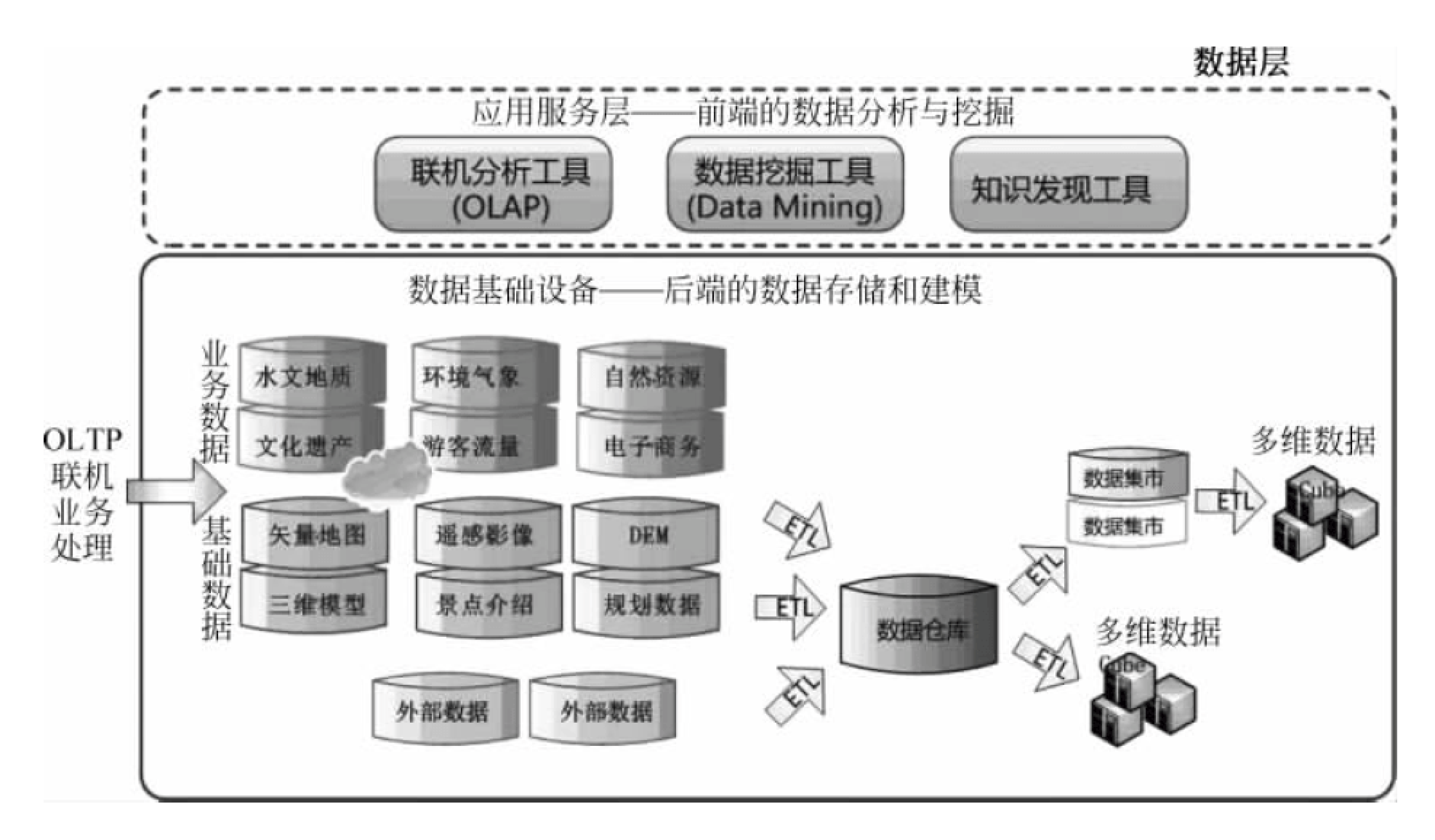


图 2-10 旅游大数据信息化平台

3. 信息管理平台

旅游景区信息管理平台是最重要的核心平台，要能实现资源监测、运营管理、游客服务等功能，包括：

（1）地理信息系统（GIS）。同时将多媒体技术、数字图像处理、网络远程传输、卫星定位导航技术和遥感技术有机地整合到一个平台上。

（2）旅游电子商务平台和电子门禁系统。

（3）景区门户网站和办公自动化系统。

（4）高峰期游客分流系统。高峰期游客分流系统可以均衡游客分布，缓解交通拥堵，减少环境压力，确保游客的游览质量。景区可以通过预定分流、门禁分流和交通工具实现三级分流，这其中要采用 RFID、全球定位、北斗导航等技术时时感知游客的分布、交通工具的位置和各景点游客容量，并借助分流调度模型对游客进行实时分流。

（5）其他配套系统。包括规划管理系统、资源管理系统、环境监测系统、智能监控系统、LED 信息发布系统、多媒体展示系统、网络营销系统和危机管理系统等。

第二篇　大数据，一种经济资源

全球每天以EB为单位所产生的多元化数据，正在成为一种巨大的经济资源，将带来全新的创新、创业方向、商业模式和投资机会。正像阿里巴巴董事局主席马云所指出的：我们正在经历从IT（信息技术）时代到DT（数据技术）时代的发展，数据的价值被逐渐发现，未来制造业的最大能源不是石油，而是数据。

从数据到经济资源，还需要实现对数据中的隐私保护，才能够成为可用数据。从经济资源到创造绝对的经济价值，则需要实现数据的标准化并提升数据质量，从而实现数据的开放和共享。如果把大数据比做“新型石油”，我们还处于石油的勘探和开采期。



数据价值与隐私博弈

随着新一代信息技术的发展，数据时代正在来临，我们周围的一切正在被数据定义。数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。在与我们切身相关的衣、食、住、行等方面，大数据对生活方式的改变显而易见，每天人们都会通过使用计算机、手机、GPS 等设备产生以十亿计的海量信息，这些相互作用的信息从根本上改变着世界原本的面貌。据统计，目前大数据所形成的市场规模为 530 亿美元左右。在数据大爆炸下，如何挖掘这些数据，也面临着技术与商业的双重挑战。

对消费者、公司和政府机构来说，数据安全与隐私保护问题由来已久。互联网、物联网技术的发展使数据的获取、传输、共享更加便利，数据总量也以指数倍增加。人们在离不开互联网、离不开个性定制服务、离不开 Facebook（脸书）和微信的同时，对数据隐私的安全疑问和要求则越来越严重。

3.1 数据的经济属性

每个经济社会都面临着很多的经济问题，西方经济学认为，任何社会，不论它依循何种社会经济制度进行运作，也不论它处在什么年代，其经济问题都源于一个基本的经济事实或矛盾，即人类欲望的无限性和满足欲望的资源的稀缺性，这一矛盾，也是经济学产生的根源。资源具有有限性和稀缺性，资源的自然丰度、时空分布等资源禀赋的差异以及资源的供求关系，共同决定着资源的价值。美国经济学家汤姆·泰坦伯格（Tom Tietenberg）指出“资源稀缺性越强，资源的价值量就越大。”①

所谓经济资源，必然具备有用性和稀缺性，有用性是资源之所以为资源的依据，稀缺性是经济资源之所以为经济资源的前提，而能否认识和利用这种稀缺的有用性则尚须依赖于一定的知识、技术和经济条件，因此经济资源通常被定义为具有稀缺性且能带来效用的财富，是人类社会经济体系中各种经济物品的总称。

3.1.1 经济物品与经济资源

1. 经济物品

人类的生存与发展，离不开形形色色的物品，广袤的土地，充裕的阳光以及作为生命之

① 汤姆·泰坦伯格. 环境与自然资源经济学[M]. 北京：经济科学出版社，2003.

源的水，都是大自然赋予人类的基本生存条件。同时人类也在不断生产出各式各样的物品，既有能够满足人类各种需要的物品，如房子、汽车、衣服、食品等，也在制造着为人类带来各种麻烦的物品，如垃圾、汽车尾气等。从经济学的角度来看，可以把与人类生存和发展息息相关的物品分为三类，即自由物品、经济物品和有害物品，如表 3-1 所示。

表 3-1 物品分类

物品	经济学说明	特 点	举 例
自由物品	不需要付出任何代价就能够得到的有用物品	“取之不尽，用之不竭”，对人类有用而且价格为零，因此通常不存在供求双方形成的买卖市场	阳光、空气等
经济物品	指人类必须付出相应代价才能够得到的有用物品	有用性；人们获得时通常都要花费代价；经济物品相对于自由物品来说一定是稀缺的；价格为正	房子、粮食等；各类物质、能源、信息等
有害物品	指人类必须付出相应代价才能够消除的物品	价格为负；卖者需要向买者付费	垃圾、尘垢和汽车尾气等

经济物品也叫稀缺物品，是指人类必须付出相应代价才能够得到的有用物品，如房子、汽车、粮食等，社会各行各业利用频谱的各种设备、服务也是经济物品，例如卫星、雷达、通信服务等。这类物品的共同特点是有一定的市场价格，即必须是借助生产资源通过人类加工出来的物品。

经济物品的第一个特点是有用性。第二个特点是人们获得时通常都要花费代价。第三个特点是经济物品相对于自由物品来说一定是稀缺的。人的欲望是无限的，但是人类面临的经济物品是有限的，这两者构成了一对矛盾，即经济物品的稀缺性。经济物品的第四个特点是价格为正，价格为正的物品在市场交易时，通常是买方向卖方支付代价。

2. 经济资源

经济资源又叫稀缺资源或生产要素，是指那些用于生产商品或提供服务所必须投入的物品。一般而言，人类在物质产品的生产活动中，都需要投入必要种类与数量的经济资源，例如在汽车的生产过程中，需要投入资本设备、人的劳动、土地等自然资源、企业家的经营管理才能、知识和技术、公共产品等，这些经济资源或生产要素通常都是不可或缺的。

（1）自然资源。如土地、矿藏、森林、陆地、海洋、河流等，总体上即指大自然赋予人类的一切有助于进行生产活动的自然条件。但必须指出，被改造过的沙漠和荒山不属于经济学中所说的自然资源的范畴，因为投入的劳动和资本已经改变了它的自然条件。

（2）资本。马克思说“资本是能够带来剩余价值的价值”，这个定义非常准确。资本包括物质资本和金融资本，前者如机器、厂房等，后者如有价证券、货币等。也可以划分为有形资本和无形资本，前者如机器、厂房、有价证券和货币等，后者如专利、商标和人力资源等。

（3）劳动。是指在生产过程中，人类自身所做出的贡献，是脑力劳动和体力劳动的总称。

劳动者的技能、受教育程度、事业心等因素影响劳动的质量，劳动者的劳动时间和劳动效率决定劳动的数量，所以劳动是指劳动者付出的数量和质量的有机结合。劳动是创造价

值的手段之一,因此每一个人都珍惜自己的劳动,都希望通过自己的劳动换取别人的劳动。所以,尽可能少地用自己的劳动换取尽可能多的他人劳动(市场上的产品和服务),是劳动者的习惯行为和追求的目标。在市场经济条件下,劳动者用自己的工作换来的报酬,就是市场对劳动者付出的劳动的一种估价。

(4) 企业家才能。是指企业家特有的个人素质,其作用是:组织协调其他要素进行生产;寻求和发现新的商业机会;引进新的生产技术或发明;引导和带动企业进行技术、市场和制度等方面的创新。

在现代市场经济中,企业家是企业的灵魂,是社会生产的组织者和领导者。企业家通过经营管理企业为社会创造财富,以此增进社会财富和人类福利,从而推动社会的进步。从根本上来看,企业家是研究、发现、引导并设法满足社会经济生活中经济主体对经济物品和经济资源的需求的人们。企业家通过自己的聪明才智研究、发现、引导这些需求并努力加以满足,从而为社会创造财富。特别需要说明的是经济学上所说的企业家不是一种职务,更不是一种称谓,而是代表着企业家经营管理方面的素质。

(5) 知识。也包括技术,是人类心智中积累下来的非物质财富,是人类对客观世界及其规律认识能力的总和。

人类社会自产生以来积累下来的知识存量决定了社会现在一年能够创造出多少物质财富和精神财富。知识包括社会科学知识和自然科学知识,后者中包含技术。可以说,没有知识,人类社会就无法组织生产,知识越多,人类社会在一定时期内生产的物质产品和劳务就越多。

(6) 公共产品。是指在消费上不具有排他性和竞争性的产品,例如路灯、灯塔、法律制度等。

经济学家认为,最重要的公共产品是制度,制度也包括法律在内。公共产品一般由政府来提供。这是因为公共产品的特点决定了企业无法专门生产路灯和灯塔等公共产品,否则企业的收益无法得到保障,企业就不能长久存在。其他如国防、军队、警察、监狱等也是一种公共产品。政府提供它使得每个人可以不做亡国奴,能在良好的社会治安条件下生存,因此提高人们的生活质量。由此看来,任何社会在组织生产时,都要投入政府生产的公共产品这种生产要素,作为报酬,使用公共产品的经济主体就应该向政府纳税。

以上6种经济资源是人类社会组织生产的物质基础,也是现代市场经济运行的基本保证。6种资源互相配合,共同发挥作用,生产出各种产品满足社会的需求。

3. 经济资源与经济物品的关系

经济资源和经济物品在语义上有着明显的区别。在谈论经济资源的时候,往往是将其作为某一类经济物品的集合,每一种经济资源实际上又可能包含成千上万种经济物品,例如,石油和石英是不同的经济物品,但都属于矿产资源或物质资源或自然资源的范畴。经济资源通常被定义为具有稀缺性且能带来效用的财富,是人类社会经济体系中各种经济物品的总称。

进一步研究经济资源与经济物品的关系,可以得出:尚未进入人类社会经济体系的经济资源只是自然界体系中的客观存在,只是经济资源的潜在形态,或潜在的经济资源,例如,未经人类开发的土地、频谱、信息等。只有在一定的知识、技术和经济条件下,潜在的经济资源才有可能被纳入人类社会经济体系之中,变为现实的经济资源。经济资源并非指向经济

物品本身，而是经济物品中相对于人类社会而言的有用性，亦即人类社会经济体系在一定的知识、技术和经济条件下，根据自身需要而开发出来的附着于物品之上的使用价值，使用价值同其物质载体密不可分，经济物品只是经济资源的物质载体，一种物品可承载多种资源功能，一种资源功能可由多种物品所承载。

3.1.2 数据信息转化为经济资源

数据本身不具备物质的实体，也不具有独立的经济价值，但是数据是信息的载体，信息是有背景的数据，经过人类的归纳和整理，最终呈现规律的信息则是“知识”。因而信息资源可以被理解为用以指引人类社会经济活动的载信物质或载信能量。大数据可以被看做依靠信息技术支持的信息群。

从社会宏观角度根据数据的产生主体可以将数据概括分为三类。

1. 政府数据

各级政府各个机构拥有海量的原始数据，构成社会发展与运行的基础，包括形形色色的环保、气象、电力等生活数据，道路交通、自来水、住房等公共数据，安全、海关、旅游等管理数据，教育、医疗、信用及金融等服务数据。在具体的政府单一部门里面无数数据固化而没有产生任何价值，如果让这些数据流动起来，综合分析并有效管理，将产生巨大的社会价值和经济效益。

2. 企业数据

企业离不开数据支持有效决策，只有通过数据才能快速发展，实现利润，维护客户，传递价值，支撑规模，增加影响，撬动杠杆，带来差异，服务买家，提高质量，节省成本，扩大吸引，打败对手，开拓市场。企业需要大数据的帮助才能对快速膨胀的消费者群体提供差异化的产品或服务，实现精准营销。网络企业应该依靠大数据实现服务升级与方向转型，传统企业面临无处不在的互联网压力，同样必须谋求变革，实现融合不断前进。

3. 个人数据

每个人都能通过互联网建立属于自己的信息中心，积累、记录、采集、储存个人的一切大数据信息。根据相关法律规定，经过本人亲自授权，所有个人相关信息将转化为有价值的数据，被第三方采集，可以快速处理，获得个性化的数据服务。通过信息技术使得各种可穿戴设备，包括植入的各种芯片都可以通过感知技术获得个人的大数据，包括但不限于体温、心率、视力各类身体数据以及社会关系、地理位置、购物活动等各类社会数据。个人可以选择将身体数据授权提供给医疗服务机构，以便监测出当前的身体状况，制定私人健康计划；还能把个人金融数据授权给专业的金融理财机构，以便制定相应的理财规划并预测收益。当然国家有关部门还会在法律范围内经过严格程序进行预防监控，实时监控公共安全，预防犯罪。

数据经济价值的形成过程中，需要将具有经济属性的经济物品，即原始数据通过技术和分析的开发转变为可创造经济价值的经济资源，即信息资源。劳动力、资本、技术等开发投入与数据资源可转化的经济价值正相关，也是数据资源产生经济价值发挥经济效益的驱动要素。

如图 3-1 所示为数据转化为经济资源的过程。

其中个人的大数据严格受到法律保护，其他第三方机构必须按法律规定授权使用，数据

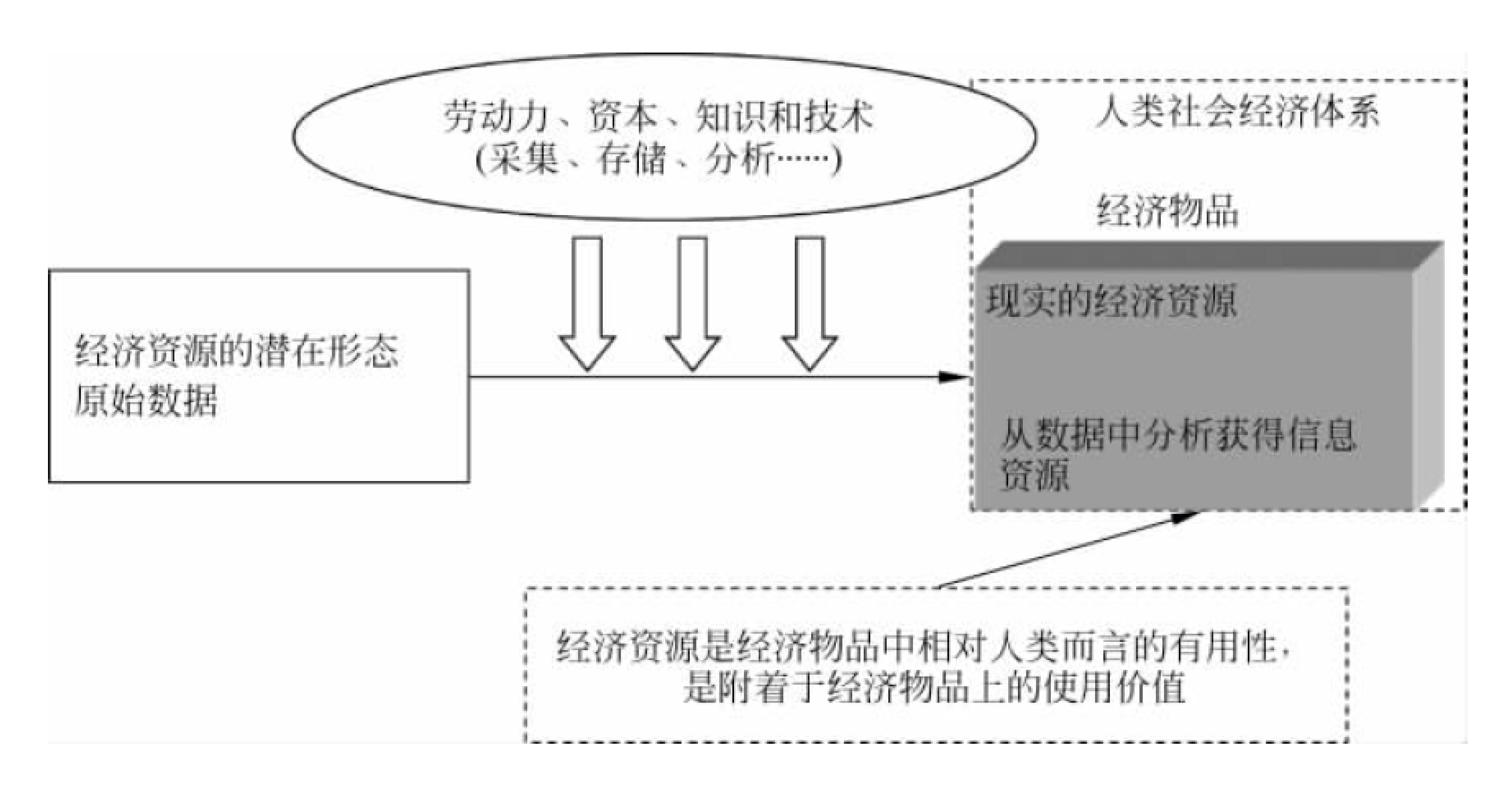


图 3-1 数据转化为经济资源的过程

必须接受公开透明全面监管；采集个人数据应该明确按照国家立法要求，由用户自己决定采集内容与范围；数据只能由用户明确授权才能严格处理。

同时，个人所产生的数据种类最为丰富，其商业价值高但密度低的特征也最为明显，也是当前通过广告投放领域实现大数据变现的主要数据来源。大量个体所产生的涉及隐私的数据之所以受到关注，不在于个人隐私本身，而在于数据本身所具有的经济资源价值。不论是数据的获得方还是应用方，获取、存储、分析数据的目的不在于挖掘个体的隐私信息，而是通过个体的多维数据重构消费者立体模型。如果从个人数据维度和群体数据维度来分析，数据的主要应用体现如图 3-2 所示。

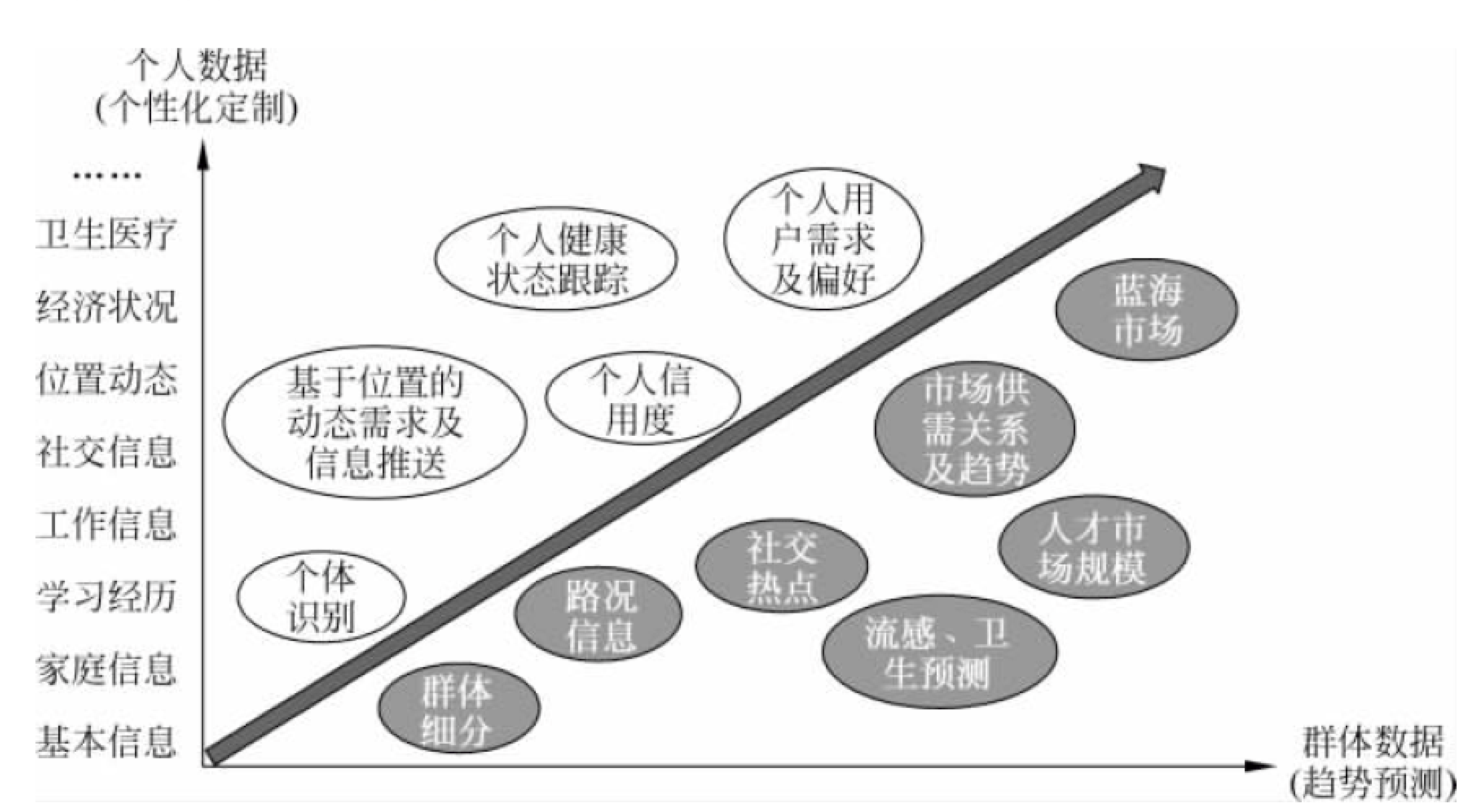


图 3-2 个人数据的应用维度

整体来看，大数据应用尚处于从热点行业领域向传统领域渗透的阶段。中国信息通信研究院的调查显示，大数据应用水平较高的行业主要分布在互联网、电信、金融行业，一些传统行业的大数据应用发展较为缓慢，批发零售业甚至有超过 80%的企业并没有大数据应用计划，远低于整体平均水平①。

① CAICT 中国信通院. 大数据白皮书(2016 年)[R]. 2016.

3.2 大数据时代的个人隐私

2013 年 6 月，英国《卫报》和美国《华盛顿邮报》爆出的“棱镜门事件”，指美国国家安全局(NSA)和联邦调查局(FBI)于 2007 年启动了一个代号为“棱镜”的秘密监控项目，直接进入美国网际网路公司的中心服务器里挖掘数据、收集情报，包括微软、雅虎、谷歌、苹果等在内的 9 家国际网络巨头皆参与其中。“棱镜”监控的主要信息有 10 类：电邮、即时消息、视频、照片、存储数据、语音聊天、文件传输、视频会议、登录时间和社交网络资料，这些数据细节使得 NSA 几乎可以实时监控一个人正在进行的所有网络搜索内容、位置信息、社交圈、消费记录等①。

“棱镜门”事件加剧了人们对大数据安全与隐私的担忧，我们所有人都在以一种前所未有的速度衍生新的数据，例如在中国，人们在网络媒体社会每分钟都产生着大量的各类数据，如图 3-3 所示。



图 3-3 中国社会化媒体表现

(数据来源：199IT 互联网数据中心②)

① 揭秘：棱镜计划[OL]. 凤凰网资讯，2013. http://news.ifeng.com/world/special/sndxiemi/content-4/detail_2013_06/13/26366771_0.shtml.

② CIC：60 秒看中国社会化媒体表现——信息图[OL]. 199IT 互联网数据中心，2016. http://www.199it.com/archives/537121.html.

人们在互联网上的一言一行,网站都会通过 Cookie 来搜索并获取相关的浏览数据。Cookie 是在 HTTP 下,由 Web 服务器保存在用户浏览器(客户端)上的小文本文件,它可以包含有关用户的信息。淘宝知道用户的购物习惯、腾讯知道用户的好友联络情况、百度知道用户的检索习惯和兴趣点等。

在 100TB 量级以上的大数据时代,信息安全与传统安全相比,不仅是一个技术问题,也是一个社会学问题,大数据的隐私保护与安全变得更加复杂,也面临更多挑战。在大数据环境下既要确保信息共享的安全性,同时为用户提供更为精细的数据共享安全控制策略等问题也值得深入研究。

这里存在着一个与人们一般认为的因果关系相反的事实:数据时代引发各类隐私安全隐患的根本原因不是隐私本身,而是包含隐私的各类数据所蕴含的经济价值。以 BAT(百度公司 Baidu、阿里巴巴集团 Alibaba、腾讯公司 Tencent)为代表的互联网企业在建立大数据平台的过程中,采集和分析的对象不是某个人的隐私信息,而是各类信息一同描绘出的一种趋势,可以是精准定位的个体用户需求,也可以是宏观市场的趋势预测。因此,大数据时代隐私风险的实质是数据自身价值与个人隐私保护的博弈决策。

3.2.1 隐私的数据化

在互联网时代,如果把数据作为一种经济资源,个体作为数据的主要生产者之一,所产生的大数据中包含大量的个人隐私,并且这些由大量个人产生的数据也是商家和企业最为关注的、具有商业价值的数据。举例来说,一个标准的美国上班族每年可以贡献 180 万 MB 的数据量,平均每天则有约 5000MB,这其中包括下载的电影、文档、电邮以及这些数据通过移动或非移动互联网传播时所产生的附加数据量,个人隐私就分散在数据化的个人信息中,如图 3-4 所示。



图 3-4 用户在线行为数据全记录

以个人用户的在线行为为例,个人浏览网页、搜索关键词、位置信息、照片分享等行为数据都会以 Cookie 的形式被全部记录下来,所涉及的个人信息和偏好隐私也将蕴含其中。这些数据中"你"的信息被保存在各种数据库中,随时可以被商家拿来识别"你"。个人的隐私分散在无数个微小的数据单元之中,这就好像一套积木,一块积木不能给予我们什么信息,但是随着积木块数的增加,整体的图景也就清晰了。

目前,各国(地区)理论和立法中"数据"与"信息"两个概念交互使用。具体而言,在个人信息保护立法中,欧盟及其成员国大多以"数据"来表述其立法保护对象。如欧盟议会于 1995 年通过的《数据保护指令》(*The data protection directive*),英国的《1998 年数据保护法》(*Date Protection Act* 1998)等①。

① 梅夏英.数据的法律属性及其民法定位[J].中国社会科学,2016,(9):164-183.

1995 年，欧盟出台的隐私法例将“个人资料”定义为可以直接或间接识别一个人的信息。很显然，当时立法者考虑的是那些带有身份标识号的文件资料之类的东西，这些标识号就好像人的姓名，而立法者们希望它们可以得到保护。

2012 年 12 月 28 日，全国人民代表大会常务委员会通过了《关于加强网络信息保护的决定》，其中规定：“国家保护能够识别公民个人身份和涉及公民个人隐私的电子信息，任何组织和个人不得窃取或者以其他非法方式获取公民个人电子信息，不得出售或者非法向他人提供公民个人电子信息。”实施这个决定，需要在操作层面上完善与之对应的法律体系，为数据保护法律体系的完善提供理论支持。

2013 年 2 月 1 日，我国首个个人信息保护国家标准《信息安全技术　公共及商用服务信息系统个人信息保护指南》正式实施，该指南比较全面地规范了个人信息处理的全流程活动，规定了个人敏感信息在收集和利用之前须获得个人信息主体明确授权。

3.2.2　数据的商业化

大数据作为经济资源，其价值的体现不仅在于对数据本身蕴含的信息进行分析、挖掘，产生有效的统计结果或个人偏好结论，从而实现其经济资源的转化。同时，数据本身也可以被二次使用或被私下买卖，转变为直接的经济利益。

2012 年 12 月成立的中关村大数据产业联盟，提出“智库、传播、资本”三位一体的新兴科技服务业模式，并以“让天下数据便捷在线流通”为宗旨建立了数据超市[①]，为数据拥有方和需求方提供数据交易平台。截止到 2016 年 9 月 2 日，完成了 928 次数据流通，建立了 435 个数据项目，数据内容近 60 万。

如图 3-5 所示为最受关注的流通数据。



图 3-5　最受关注的流通数据

大数据时代，人们加快脚步将海量用户数据转换成有用的知识并揭示其潜在价值，这个潜在价值更多的时候通过对数据的再利用、不同数据的重组、数据扩展等来开发二次使用价值。在隐私数据化后，这个过程中必然更多涉及对人们行为习惯、消费习惯、个人喜好等个人隐私的挖掘，个人隐私变换成现实的商业利益。在信息时代，个人的隐私变得有利可图，

① https://hub.dataos.io/.

从而使得隐私变成可以买卖的商品,隐私也就具有了价值的特点[①]。

3.2.3　个性化服务的博弈

每个人都期待获得个性化服务,当我们浏览淘宝网后再去打开其他的网页,会看到当前页面推送了我们刚才在淘宝上所看到的商品的广告,这是网站上的 Cookie 对我们行为进行跟踪的结果。但是在大数据时代,想要获得个性化服务,就一定会在某种程度上牺牲自己的隐私。

与大家的常规想法不同,当事情涉及广告和隐私时,广告商并不在意我们在做什么或我们想要什么,他们只关心一件事情:让我们去买他们销售的东西。你可能想问:那又怎么了,谁不喜欢在寻找一件特定商品时正好收到相关广告和讯息,谁不想要在临近最喜爱的商店时收到他们的电子优惠券呢?

这似乎是个不错的交易:只要舍弃一点儿私人信息,就可以得到一些帮你省钱的、有用的免费服务。在这个看似双赢的局面下,实际上是用户用自己的隐私换来个性化服务并给商家带来巨大的利润。被我们“免费”分享的信息不只是被广告商用来销售商品给我们的,这些信息也被 Facebook 和谷歌(Google)这类公司以及其他各种商家用来分析、预测用户行为,从而发家致富。

当你在使用电子邮件、社交网络的时候,你大概也会知道你的信息正在被记录下来,你发表的言论或者分享的照片、视频等都决定着互联网运营商即将向你推荐什么样的资源和广告;当你拿着 iPhone 满世界跑的时候,苹果早已通过定位系统把你的全部信息搜罗在自己的数据库里,利用这些信息来构建地图和交通信息等;当你在享受着视频监控带来的安全感的同时,别忘了你也是被监控的一分子,你的一举一动都会暴露在镜头下面;你用手机通话时,运营商不仅知道你打给谁,打了多久,还知道你是在哪里进行的通话。周鸿祎指出,大数据时代可以不断采集数据,当看起来是碎片的数据汇总起来,“每个人就变成了透明人,每个人在干什么、想什么,云端全部都知道。”

大数据是好的时代,也是坏的时代:如果免费检测基因的公司拿到了个人的健康隐私数据,就能精准地推销医药产品,建立点对点的商业模式,这对公司是一个黄金时代。但如果大数据被污染了,也就是说,数据被人为操纵或注入虚假信息,据此做出的判断就会误导人们。在大数据时代,曾经在全世界范围内达成共识的“告知与许可”隐私保护政策正面临失效。

任何技术都具有两面性,大数据技术也是这样。随着企业对用户数据无限制搜集,数据交易、数据整合使得企业对个人数据的挖掘越来越深入,公民隐私泄露的问题也变得越来越不能被忽视。关于个人信息保护的立法工作一直处于进行阶段,但还没有跟上时代的发展变化。虽然在大数据时代人们的隐私更容易暴露,但是不能因此排斥大数据,而是应该在数据挖掘和公民隐私之间找到一个平衡点,最大限度地利用大数据技术来便利人们的生活。

---

① 黄成鹏. 大数据时代的个人隐私[D]. 华中师范大学,2015.

3.3 旅游大数据应用价值

随着互联网、移动互联网、物联网的发展，数据蕴含的价值也在增加。针对数据价值的利用，可以简单地分为以下三个层次。

（1）数据查询。从海量的数据中快速定位到目标信息。

（2）数据统计。从海量的数据中根据不同的维度和颗粒度快速地生成统计信息。

（3）数据挖掘。从海量的数据中发现规律和关联关系来辅助决策。

三个层次层层递进，其实也是对数据利用的不断细化和深入。旅游市场规模的扩大也吸引了大量旅游相关主体参与享受红利，旅游企业之间的竞争的日益激烈，传统旅游实体的经营方式，往往由于过度依赖自身的资源而忽略消费者本身的旅游需求，经营具有盲目性、经验性。在移动互联网时代大背景下，游客获得信息的手段和效率远远优于以往，旅游企业、景区之间如何在新战线上提高自身品牌影响力、如何吸引更多游客、增强游客的购买性、提升游客的旅游体验成为新常态背景下的旅游行业。

3.3.1 数据推动旅游行业价值流动

旅游行业是高度依赖“信息”的产业，正是“信息”造成了旅游行业的价值流动。如果说物联网与云计算等信息技术正在重构现代旅游行业，那么“大数据”则是这次产业升级的关键。大数据作为新兴的技术手段，为解决以上问题提供了新思路、新路径，大数据应用主要集中于游客洞察与服务、景区管理两个方面。

1. 游客洞察与服务

全面深入地了解游客或潜在目标客户，是旅游景区做好景区产品营销、提升游客体验的重要前提。旅游时游客在“吃、住、行、游、购、娱”一系列旅游行为动作的过程中，通过线上、线下等各种渠道留下游客行为痕迹、碎片化数据，将这些数据通过不同路径、不同渠道进行整合汇聚，再通过大数据技术，可以从多维数据中挖掘出对旅游景区业务运营有价值的信息点，具体主要体现在游客来源分析、游客口碑分析、游客特征分析、游客行动轨迹分析等方面。

基于对游客的认知，旅游景区可以有针对性地进行线上、线下营销推广，并根据游客偏好配置景区相关的资源，提升景区的服务水平，实现“智慧服务”。

2. 旅游景区管理

旅游景区有着地域面积广、管理要素多等特点，传统的管理方法存在景区无法全面覆盖、管理效率跟不上旅游市场的快速增长等问题。借助物联网技术，利用视频、GPS等技术手段搜集各类数据，并依托大数据平台进行管理，完成旅游景区各类要素的数字化、可视化管理。同时，产生的各类运营数据，经过分析，指导旅游景区的建设和优化，实现了基于数据的“智慧管理”。

大数据不仅是一项技术手段，更是体现量化管理的一种思维方式，其应用的思路包括数据从哪来和数据怎么用两个基本问题。在实际的大数据应用过程中，瓶颈也往往在于“有需求没数据”和“有数据没需求”两个基本矛盾点，无法实现从基础数据到业务应用的价值链闭环。

旅游景区“智慧管理”主要体现在客流量监控预警及分析、GPS 定位数据跟踪分析、森林防火智能检测等方面。本案例以某景区大数据平台建设为背景，着重阐述“数据汇聚”和“数据应用”两个方面，其中，“数据汇聚”介绍从数据源以及基础平台承载的方面，“数据应用”侧重于旅游景区“智慧服务”和“智慧管理”两个应用链条的实现思路。

数据应用是数据价值释放的最终环节，也是最重要的环节，明确数据应用的方向决定了整体大数据解决方案的实施框架和思路，是大数据平台建设的前提条件。目前，大数据应用比较成熟的领域集中在金融、电信、互联网、零售等行业，业务应用方向集中于营销、征信等方面，对于信息化程度相对滞后的旅游行业，大数据应用更多地继承成熟行业应用方向，即基于对“人”的洞察实现多类型的应用。同时，旅游行业的特点使得其对“物”的管理要求比较高。

基于大数据的智慧旅游解决方案，应用层面集中于“人的洞察”和“物的管理”两个层面，对应到本次案例的“智能服务”和“智慧管理”。

3.3.2　智能服务

旅游作为服务行业，服务品质是旅游景区的核心竞争力之一，大数据分析作为技术手段，可以通过洞察了解游客特征，优化服务策略，也可以整合多类数据，开发数据产品，为游客提供信息服务。

1. 游客来源分析

旅游行业的市场竞争往往是全国性的、区域性的，如何将有限的营销宣传成本有效地转化置换为客流是旅游景区急需解决的课题。“客户在哪？他们是谁？有哪些需要特征？”是做好宣传推广需要回答的关键问题。互联网和电信网络具有全域覆盖的属性，通过大数据技术采集景区相关的数据，可以洞察到线上搜索以及线下到访中各个区域的人数规模，指导旅游景区安排线下资源进行定点投放，针对性的提高使得营销成本利用更为高效，客源地游客转化效果也相对更好。

通过与外部数据资源厂商建立合作，获得全国范围内各类平台的潜在游客数据，其中主要从互联网数据中提取到 LBS 用户定位、搜索数据、电商数据、社交数据等，获得游客的旅行意向轨迹。与电信运营商合作，分析到访游客的区域归属信息，追溯游客的旅行轨迹。

整合线上潜在游客归属和到访游客的归属数据，可以构建出旅游景区的游客归属地分布地图，并以此探索市场洼地。掌握了客源分布的同时，综合客源地的经济发展情况进行宣传推广的优先级区分，从开放的渠道获得区域经济发展水平、人均 GDP 等数据，进行客源地价值细分、评估消费能力及对旅游景区所能带来的潜在营业额，在市场资源投放时做到有的放矢，精准的宣传可以大幅提升客源市场的转化率，实现降本增效。

2. 旅游景区口碑分析

依托强大社交网络和资讯获得渠道，单个游客在线上所发表的言论，在条件适当的情况下就会变成热点事件，在短时间就会有足够大的覆盖面，比如“青岛天价虾”“哈尔滨天价鱼”等事件就是通过社交媒体的传播，对当地旅游造成了极负面的影响。

良好的口碑是景区构建品牌、确保客流的重要保障，通过获取全网数据，包括论坛、贴吧、微博、新闻等网站数据，采用网络文本挖掘技术，实时监控旅游舆情，及时发现游客负面

反馈，消除或改善不安全或游客不满意的项目和产品。景区的口碑分析包括以下分析维度。

（1）正负面评价。整合线上与景区相关的数据，通过文本分析，识别游客在景区相关描述中的情绪的正负属性，根据正负比例情况判断景区网络口碑。

（2）游客负面反馈监控。通过线上数据文本挖掘，识别出游客主要的负面反馈关键字，及时地、有针对性地改善景区服务。

（3）游客景区评价分析。建立景区设施、交通、服务、价格、卫生、餐饮等6个主要维度，通过景区自建的手机服务平台和线上数据分析，获得游客在各个维度下的评价数据，从而有针对性地进行提高。

（4）网络关注度。根据搜索数据、媒体报道等，获得旅游区域热点信息、同区域内地热门旅游景点信息，指导景区的宣传推广工作。

（5）舆情监控。监控各主流媒体和旅游网站相关的资讯报道，评估各个报道的影响力，做到对景区负面消息的及时监控和响应。

3. 游客特征分析

根据游客的诉求开发旅游产品、设计旅游服务，是为游客提供优质体验的关键，在这一环节，尽可能多地了解游客是重要前提，游客特征分析的目标就是借助大数据，更多地了解游客各方面的信息，并把这些信息标签化，支撑景区运营。

游客特征分析，一方面通过分析到访游客的结构和特征，掌握旅游景区的核心客户群体的特征，并以此为基础优化景区的服务能力；另一方面，通过抓取线上与景区相关的数据，锁定景区的潜在目标客户群，再通过与第三方宣传渠道合作，实现景区的精准营销。

旅游景区引入电信运营商和互联网平台等丰富的第三方数据资源，构建全面洞察线上、线下游客特征的标签体系，从多个维度洞察包括身份属性、消费能力、行为特征、偏好在内的多类信息。

1）到访游客分析

通过与电信运营商合作，得到到访游客的特征分析，包括性别比例、年龄分布、身份画像等多类个体属性标签，也包括游客的驻留时长、驻留主要区域等信息，在数据资源合作的过程中，所有的数据均通过统计数字得到展现，并不锁定到某一个独立的个体，规避了个人隐私泄露的风险，掌握主要到访游客的特征，可以有针对性地优化景区的产品和服务，提升游客体验。

2）线上潜在目标客户分析

与互联网数据服务提供商合作，跟踪各类与旅游产品相关的网站，从第三方DMP平台获得特定终端用户的浏览数据，通过搜索关键字和浏览的页面信息，掌握该终端用户对旅游景区、旅游产品、价格等多方面的偏好数据，对契合度较高的潜在客户群体，进行定向的景区广告投送，提高线下游客的转化率。

4. 景区商户评价

考虑到旅游行业的特殊性，游客在景区内的商户消费属于低频交易行为，对于商户而言，其收入结构更多依赖于新增游客的一次性消费，存量消费占比很低，这样的特征导致全国各地的景区商户频频上演“杀客”的现象。商户是旅游景区的重要组成部分，与景区在对外口碑宣传方面是利益共同体，景区内或景区周边商户的服务能力和质量，会影响游客对景

区的口碑和旅游体验。

大数据技术使得景区具备了从多个角度监管商户的能力，数据的获取渠道有景区的投诉热线、景区手机服务 App 平台、开放的互联网平台、景区为商户配置的 POS 机、监控摄像头等。

其中，互联网平台抓取与商户有关评论数据，通过文本分析抓取游客反馈的关键字，在掌握游客对商户正面负面评价比例数据的同时，还可以从评价文本中提炼游客反映的意见，有助于进行定点的督导改进。通过 POS 机、监控摄像机能够全面掌控商户实时交易数据，为商业坪效（单位面积单位时间内的销售产出）评价提供数据支持，对比商业营业额与周边客流量变化、不同消费项目的消费额等，为商业数量增减、商户经营品类调整提供依据。

在利用技术手段实现监控和监管的同时，景区针对覆盖范围内的商户建立长期有效的服务评级机制和黑名单制度，把商户评价评级的主导权交给游客，依托景区手机服务 App 实现游客的线上评分和投诉，助力景区商户精细化管理。

5. 游客手机服务平台

智能手机的高普及率使得手机成为旅游景区提供服务的优质触点平台，大数据在手机服务平台上的应用更多地体现在整合多类数据，开发数据产品，并以 App 的方式为游客提供信息服务上。

（1）电子门票：依托手机服务平台，以二维码电子门票替换原有的纸质门票，通过手机平台实现门票购买、在线支付、订单查询、订单管理、退票和检票功能，进入景区可扫描二维码或者身份证，提升了游客进入景区的体验。

（2）智能导览：智能导览是利用二维码技术、GPS 定位技术、文本转语音技术、电子地图、多语言支持等技术相结合，为游客提供自助导览服务。通过智能导览服务，游客可以实时掌握在景区的位置，并获得推荐游览路线，当游客进入景点位置 5m 范围内时，服务平台会提示并提供景点语音介绍服务。

（3）景区商户推荐：手机服务平台会推荐给游客住宿、餐饮和购物的商户信息，支持游客对商户的服务评价和星级评级，评价和评级的历史数据会沉淀在平台供游客进行比较选择，并支持游客的线上预订服务。

（4）一键呼救：游客使用一键呼救功能，系统自动拨打调度中心报警电话，同时自动打开终端 GPS 芯片进行定位，将游客的位置信息发送到调度中心 GIS 平台进行定位，方便监控中心进行警员调动、警员接警后的快速反应，最大程度保障游客利益。

（5）常用信息服务：提供景区介绍、天气、交通、旅游指南、公厕位置、银行等配套设施位置等多类信息服务。

3.3.3　智慧管理

景区的智慧管理体现在对景区内游客状态的实时监控、游客量的预测以及大量的自然资源、人文景观、游乐场所等固定设施的监控。通过监控，能够预防和快速处理突发事件的发生。同时，沉淀的数据，如客流轨迹也有助于优化景区内各类资源的配置。

1. 客流量监控预警及分析

智慧景区监控系统是大数据平台的重要基础设施，可对突发事件如踩踏、拥挤进行实时

监测，及时预警。本案例中的客流量监控预警系统，建立了对景区主要道路及路口的实现客流量计数、流量分析、系统示警和流量调控等功能的智慧化综合管理平台。对景区现场实施全天候、全方位 24 小时监控及人员流动的记录，达到加强现场监督和安全管理的目标。

根据景区客流系统的整体规划，在本次客流统计系统中各重点出入口、重点路段将作为最基础的客流采集单位，客流数据经过内部网络传回至景区信息监控中心。本案例采用国际领先技术的、高精确度的 Smartcount 客流统计产品。Smartcount 采用独有的立体图像处理技术和追踪追尾处理，用一台摄像机进行三次元处理，可以实现更高精度的数据采集。同时采用人形立体识别技术，可识别人的立体形状，实现高速高密度分析处理，可在计数范围内从各个角度把握、追踪人像，很复杂的人的行为也可准确计数，同时采用独有的店员排除计数的算法，可更精确地掌握实际客流量。

系统的视频显示功能形成实时的景区热力图（如图 3-6 所示），最大程度地帮助景区管理人员了解和分析园区的客流变化。除了实时的客流展示，还能够提供选定时段客流量趋势变化、客流数量对比、人均停留时间及对比、人均停留时间趋势变化等维度的指标分析，支撑景区内部各个层级管理人员的人流量监测需求，满足政府对整体旅游市场人流量的统计和监管需求。



图 3-6　景区实时热力图

2. 游客流量预测

国内旅游市场存在明显的潮汐现象，每年旅游旺季都会有个别景区出现客流量“井喷”、旅客滞留的现象，如果在旅游旺季到来之前，景区能够主动探寻游客的消费动向，较为准确地预测客流量，提前准备相应的应对方案、配置资源，能够一定程度地解决因游客流量暴增导致的交通、住宿、安全等一些问题，提前引流或处置资源，保障游客在景区内的旅游质量。

大数据平台综合分析景区积累的历史数据，分析趋势，确定需要重点预测的时间窗口，与各大旅游网络平台和互联网公司合作，获得机票、酒店、度假、门票、景区搜索量等外部数据，汇总、分析各个数据源的数据结果，可以提前数日预测各景区到访人数。

线上的游客人流量预测可以预估景区的热度，得到一些粗粒度的统计数据，此类数据可以在宏观层面指导景区配置资源，但在实际的操作过程中，线上数据的预测数据本身存在一定偏差，再加之有相当规模的临时性的散客，无法通过早期的线上痕迹进行监测，针对这种情况，景区与交通部门和电信运营商合作，在旅游旺季每天获得准实时的车流量和景区辐射范围内的手机连接数据，从侧面预测到访的游客流量，线上线下数据相互验证、补充，时间窗口预测和实时监控预测相结合，能够相对准确地预知每日到访景区的人数。

3. 景区地理信息汇集分析

旅游与地理信息相关性极强，GIS 中如图形、区域景观资源信息、交通路线等诸多要素与旅游密切相关。GIS 支撑下的旅游信息系统与一般旅游信息系统相比，可以完成一些特殊功能，如图形分析，空间数据综合处理、分析等功能。随着旅游业的迅速发展，传统的旅游地图已远不能满足人们的需要。以空间信息处理为核心的地理信息系统技术，因具有强大

的空间信息管理、空间信息分析、空间信息查询及三维影像显示等功能，而成为旅游业信息化的首选平台。

旅游地理信息系统（Travel Geographic Information System，TGIS）在这样的背景下应运而生。TGIS是以旅游地理信息数据库为基础，在计算机硬软件支持下，运用系统工程和信息科学的理论和方法，综合地、动态地获取、存储、管理、分析和应用旅游地理信息的多媒体信息系统。作为旅游景区大数据平台的重要基础设施，TGIS的建设可以实现景区资源的系统化管理并支撑智慧化的旅游服务。

1）景区资源系统化管理

构建统一的信息共享平台和指挥调度体系，围绕景区的资源保护、经营管理、安全防范和可持续发展等方面的应用，实现景区信息数字化、应用网络化、服务智能化，更好地保护和开发景区旅游资源，为景区的科学管理、发展决策提供信息技术支持。

2）支撑智慧化旅游服务

建立以景区地理信息系统为平台的旅游服务体系，支持景区的旅游资源调查与评价、旅游规划、景观设计、配套服务设施建设、旅游商品设计销售、旅游资源及生态环境保护等，满足旅游服务与管理的需要。

4. GPS定位数据跟踪分析

通过景区地理信息系统，整合游客、工作人员、车辆的实时GPS定位应用，从而既可形象直观地实现对内部工作人员的日常到岗管理，又可以在发生紧急情况的时候，能快速准确地进行救援决策，调度各界力量实施科学、有效的救援工作。

GPS终端分为车载设备和手机两类，可以实现对车辆和个体人员的定位监控。GPS定位数据跟踪主要包括以下模块。

（1）员工与车辆定位监控。可以随时掌握景区内员工和车辆的分布情况，并根据需要进行调度。

（2）多点监控。依据管理权限树，按照层级不同，可以对该层级下可以监控的人员或车辆信息实现多点实施监控。

（3）单点跟踪。锁定特定监控对象，跟踪被监控对象的运行轨迹以及相关信息，如位置、速度、行驶方向等。

（4）轨迹查询。查询任意监控对象的历史运行轨迹，并可在三维地形上展示。

（5）报警。GPS终端根据报警配置情况，可实现主动报警，比如超速报警、跨区域报警等。

5. 森林防火智能监测

森林防火智能监测平台采用无缝融合智能图像识别技术、面向对象的3D GIS技术、大型网络监控技术等高新技术，结合林业管理的专业知识和林业防火的经验，建立林业防火智能监测预警及应急指挥系统，从而实现林区视频的自动监控、烟火准确识别、火点精确定位、火情蔓延趋势推演、扑救指挥的辅助决策、灾后评估等多方面功能，建立森林防火的完整业务链，并针对性地解决用户的各种个性化需求。

通过在林区高处安装三维精确定位摄像系统获得林区的清晰图像，利用视频分析技术，根据烟、火的光谱特征判断是否发生火灾。一旦发现疑似火情，立即触发报警，林区视频回

传至监控中心，如果确认报警属实，摄像系统锁定目标，精确判断火点位置，并根据已建立的林业防火信息数据资源做出灭火方案及灾后评估。

（1）防火准备辅助决策。根据各地区的火险天气预报和火险等级预报，为各地、各林区分别提供各自不同的火灾预防措施、火源管理措施、扑火队伍战备措施等辅助决策意见。

（2）林火行为预测。林火一旦发生，系统可迅速向决策者提供林火区位、蔓延速度、火场扩展趋势、火线强度等重要的火情数据。

（3）扑火辅助决策。根据林火发生地的动态信息，利用本系统的虚拟演示实现对扑火工作的复杂指挥。

（4）火情蔓延推演。动态推演火灾蔓延的方向、速度、区域等。

（5）三维应急指挥：借助电子沙盘进行复杂的应急指挥。

（6）灾后评估总结。借助 GIS 数据库，对过火面积及火灾损失进行评估，并对灾后重建提供决策依据。

旅游行业涉及游客、景区、商户等多类要素，有诸如订票、游览、交通等多个应用场景，在线上线下不同平台上产生大量的结构化数据和非结构化数据，旅游大数据解决方案从景区内部以及外部整合数据资源，历经采集、存储和清洗进行数据处理，支撑景区内部的管理、营销和游客体验提升，其中涉及物联网的技术、线上数据的抓取技术等，改变了传统的人工对旅游数据的采集、排查、分析等工作程序，改变的不仅是决策效率和洞察维度，更是重构了传统旅游行业的整个产业链条，改变了旅游行业运营方式。



大数据的开放与共享

大数据是一场彻底改变人们生活、工作和思维方式的革命，是继移动互联网、物联网、云计算后对 ICT 产业具有深远影响的一次技术变革。在工业化向信息化转型时期，信息的公开、共享与服务成为时代发展的主题，作为信息载体的数据正在成为与物质和能源同等重要的经济资源。

目前，大数据企业不断推出各式各样的大数据存储、处理、分析产品，同时，社交网络、金融、通信、政务等大数据存在的领域，也相继建设大数据平台，从平台的存储处理分析等各方面都无不体现着行业特征。但是因数据源、格式、内容等的多样性，使得大数据的应用缺乏通用性和标准化的现状，限制了大数据的开放共享，在很大程度上也阻碍了大数据的发展。

4.1　数据资源开放和共享

未来 5 年，全球数据量呈指数级增长。据国际数据公司(IDC)统计，2014 年全球数据总量为 8ZB，预计 2020 年达到 44ZB。同期，我国数据总量为 909EB，占全球数据总量的 13%。其中，媒体、互联网数据量占比为 1/3，政府部门、电信企业数据量占比为 1/3，其他的金融、教育、制造、服务业等数据量占比为 1/3。预计到 2020 年我国数据量将达到 8060EB，占全球数据总量的 18%。[①]

大数据应用的关键在于分享，各行业已逐渐意识到单一的数据无法发挥最大的效能，一个个信息孤岛无法独自实现真正的全数据分析，也就无法完整地重构用户画像，数据的缺失也会导致市场发展趋势预测的偏差，因此未来大数据的健康发展，资源的开放和共享是核心。

4.1.1　打破“信息孤岛”

信息孤岛的产生，应该说是大数据发展过程中的一个必经阶段，也是当前的发展瓶颈。我们在认识大大数据的发展历程时可以看到，需要数据采集、存储、处理技术的共同发展。

大数据飞速发展过程中，当数据作为经济资源的价值日益突出，数据信息孤岛的困境也随着出现。我国公共数据资源开放处于起步阶段，面临制度、规范、平台、数据可用性等问题与挑战，整体呈现“不愿开、不敢开、不能开、不会开”的局面。

① 2016 年大数据白皮书[R].

1. 不愿开

信息资源是独家垄断资源，开放后担心部门权力削弱、经济利益受损，并可能暴露部门业务问题。

2. 不敢开

尚缺乏保障数据开放的配套制度，缺少具有可操作性的强制性规定，与政府信息公开、保密法、档案法等相关法律法规衔接不到位，各地对政府数据开放的范围和潜在风险存在“后顾之忧”。

3. 不能开

数据基础不牢，公共部门尚未建立一套完整的数据资源采集、管理、加工和开发利用的体系，很多信息资源缺乏数字化，数据资源多头采集、重复建设、成本高昂，很多数据无人维护、不具有可持续性，数据的质量和准确性也存在问题，有哪些数据资源也不清楚。

4. 不会开

开放质量不高，可利用性差，网站的建设和维护问题，网站数据的质量问题(可机读性差，数据更新频率不高，数据互动性差等)。网站缺乏标准化，增加使用者成本。

4.1.2 全球数据的开放与共享

美国政府最先对大数据革命做出战略反应。2009 年，美国联邦政府发布《开放政府指令》，作为大数据的前奏推出了 Data.gov 公共数据开放网站。2012 年 3 月，美国联邦政府发布了《大数据研究和发展计划》，正式启动了“大数据发展计划”，宣布将投入超过两亿美元在大数据研究上；同年 5 月，联邦政府发布《数字政府战略》(*Digital Government Strategy*)，致力于为公众提供更好的“数字化”服务，围绕数据进行的一系列措施在美国政府全面推进，大数据对美国政府的影响逐步显现。

2013 年 5 月 9 日，奥巴马签署第 13642 号总统行政令，对联邦大数据管理工作提出了新的准则，提出在保护好隐私安全性与机密性的同时，将数据公开化以及可读写化纳入政府的义务范围。2014 年 5 月 1 日，美国总统行政办公室向奥巴马提交了一份名为《大数据：把握机遇，维护价值》的报告，阐述了大数据带来的机遇与挑战。报告认为，大数据技术为美国经济、人民的健康和教育、能源利用率以及包括信息安全在内的国家安全等提供了难得的机遇。同时，报告也指出了大数据为美国隐私保护、信息安全和社会发展带来了新的挑战。在这些战略框架中，基本都考虑了大数据对既有法律制度的挑战和相应对策。

欧盟专门在 2014 年发布了《数据驱动经济战略》，有望近期内成为欧盟经济单列行业，为欧盟恢复经济增长和扩大就业，做出巨大贡献。欧盟在大数据方面的活动主要涉及两方面内容：①研究数据价值链战略计划；②资助“大数据”和“开放数据”领域的研究和创新活动。

数据价值链战略计划包括开放数据、云计算、高性能计算和科学知识开放获取 4 大战略，主要原则是：高质量数据的广泛获得性，包括公共资助数据的免费获得；作为数字化单一市场的一部分，欧盟内数据的自由流动；寻求个人潜在隐私问题与其数据再利用潜力之间的适当平衡，同时赋予公民以其希望形式使用自己数据的权利。

《国务院关于印发促进大数据发展行动纲要的通知》(国发〔2015〕50 号)中，将全面实施大数据战略，提高信息资源掌控和利用能力，推动数据共享开放和开发利用，培育有国际竞

争力定位主要任务，对于大数据的资源开放和共享任务主要有以下几点。

(1) 统筹政务数据资源和社会数据资源，形成统一、开放、共享的新格局；

(2) 加快国家人口库、法人库、空间地理库和重要领域信息资源建设，推动形成全国统一的基础数据资源体系；

(3) 建立完善国家数据共享平台，推动跨部门数据共享、跨领域业务协同和跨区域制度对接；

(4) 加快建设国家政府数据统一开放平台，强化对国家公共数据资源的统筹管理，制定公共部门开放计划，稳步推进国家数据资源向社会开放。

在大数据与信息经济并发的时代，数据和信息资源便成为全世界公认的重要新型资源，尤其是在一些能源匮乏的地区，开放信息资源将会有利于为国家创造新的经济增长点，节约人财成本，夯实国家的基础建设。公共信息资源是由政府部门或单位通过信息与通信技术的不断更新，科技的进步，时代的发展积累与产生的信息资源，因为是由我国政府出资或是授权才能产生的资源，因此公共信息资源应是国家的战略资产和重要财富，这些资源如同能源一般蕴育了巨大的使用价值和信息财富，这些资源主要产生在政务部门、事业单位、市政公共企业事业单位等部门与单位，将这类公共信息资源进行开放共享有利于社会与公众的衣食住行、生产生活、发展娱乐等方面的日常的生活需求。

4.1.3　数据标准化

数据的开放和共享是加速大数据技术和应用发展的趋势，但是大数据的 5V 特性给数据的收集、处理和可视化等多方面带来了极大的困难，并且由于数据资源、内容、格式、采集技术等的多样性，以及隐私保护的需求，大数据的开放和共享也成为当前大数据的发展瓶颈。建立一套完整的大数据标准化体系是推进资源的开放和共享的必要工作，行业资源的共享和跨行业数据的开放都需要以一定的基础数据标准为借口。

当前的数据之所以难以开放共享，根本原因在于当前的数据整体系统的复杂性和标准化体系的缺失。如同铁路的钢轨不是统一标准，就不能连接到全国各地，甚至跨国通车。

大数据发展的关键要素主要有数据、技术、应用以及政策扶持，如图 4-1 所示。其中，数据源的采集和分类，存储和计算技术、行业数据对接等各个环节都缺少统一标准，难以实现开放共享、互连互通。



图 4-1　大数据发展的关键要素

目前，国际上主要研究制定大数据标准的组织，如表 4-1 所示。

表 4-1 国际大数据标准化组织

序号	标准组织和协会	个数	范 围
1	ISO/IEC JTCI SC7、SC27、SC38、ITU-T SG13	4	国际标准化组织
2	DMTF、CSA、OGF、SNIA、OCC、OASIS、TOG、ARTS、IEEE、CCIF、OCM、Cloud use case、A6、OMG、IETF、TM Forum、ATIS、ODCA、CSCC、W3C	20	国际标准化组织
3	ETSI、Eurocloud、ENISA	3	欧洲
4	GICTF、ACCA、CCF、KCSA、CSRT	5	亚洲
5	NIST	1	美洲
	合计	33	

自 2012 年开始，ITU-T、ISO/IEC、NIST、CCSA 等国内外标准研制组织相继组建工作组展开大数据研究和标准化工作，这些工作组在大数据定义、相关术语、需求等方面输出少量研究报告和标准。虽然研究成果有限，但其研究方法和方向具有重要的借鉴意义。①

国际大数据标准化现状如图 4-2 所示。

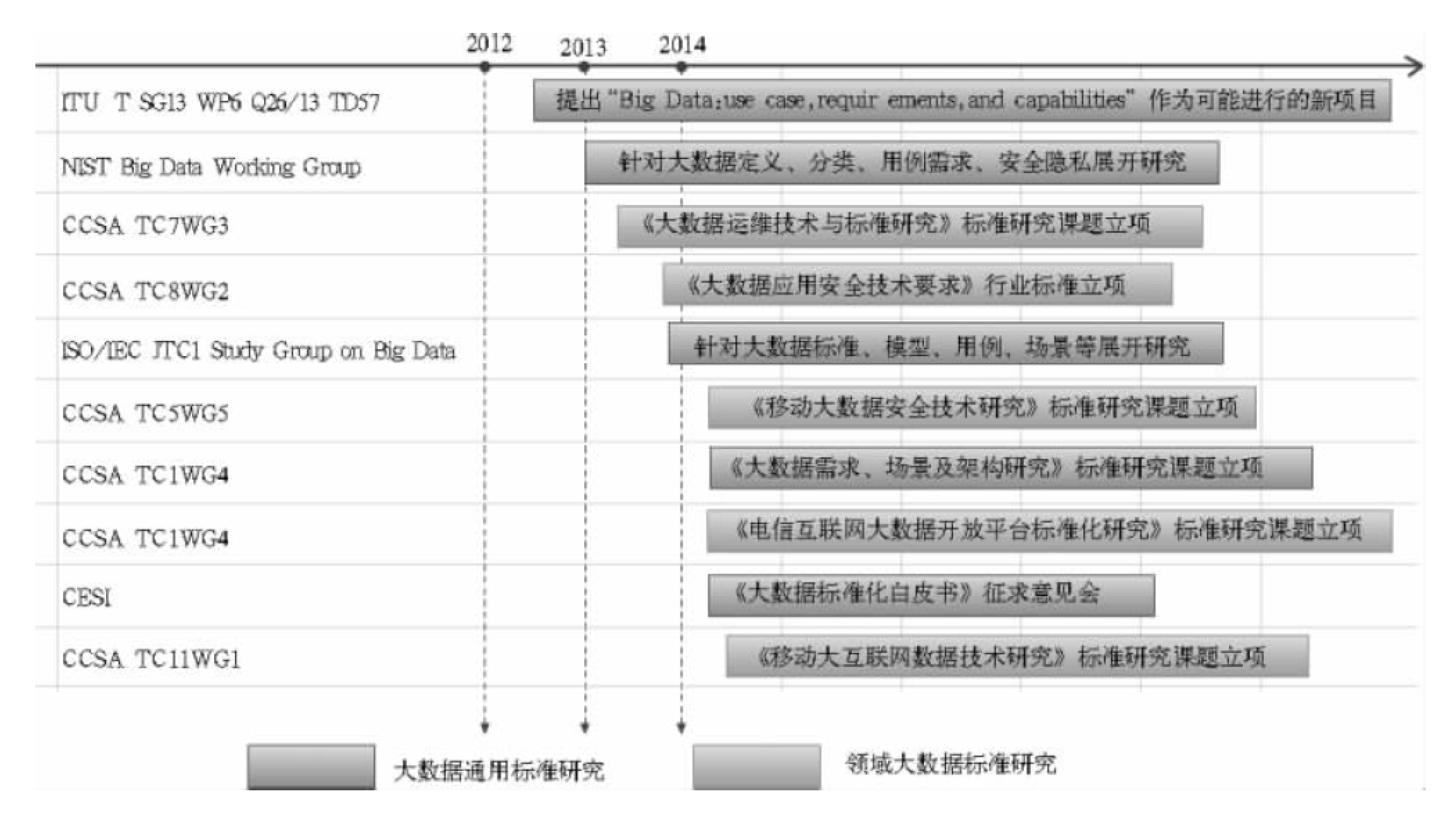


图 4-2 国际大数据标准化现状

2016 年 5 月，全国信息技术标准化技术委员会大数据标准工作组发布了《大数据标准化白皮书(2016)》，提出了由基础标准、数据标准、技术标准、平台和工具标准、管理标准、安全和隐私标准、行业应用标准共 7 个类别组成的大数据标准体系框架。其中，数据标准主要针对底层数据相关要素进行规范，包括数据资源和数据交换共享两部分。数据交换共享就是针对将大数据作为经济商品进行交易的相关标准，以及为实现政府和社会数据的开放共

① 韩晶，王健全. 大数据标准化现状及展望[J]. 信息通信技术，2014，(6)：38-42.

享制定统一的数据格式、编码、单位等标准。[①]

数据标准框架如图 4-3 所示。

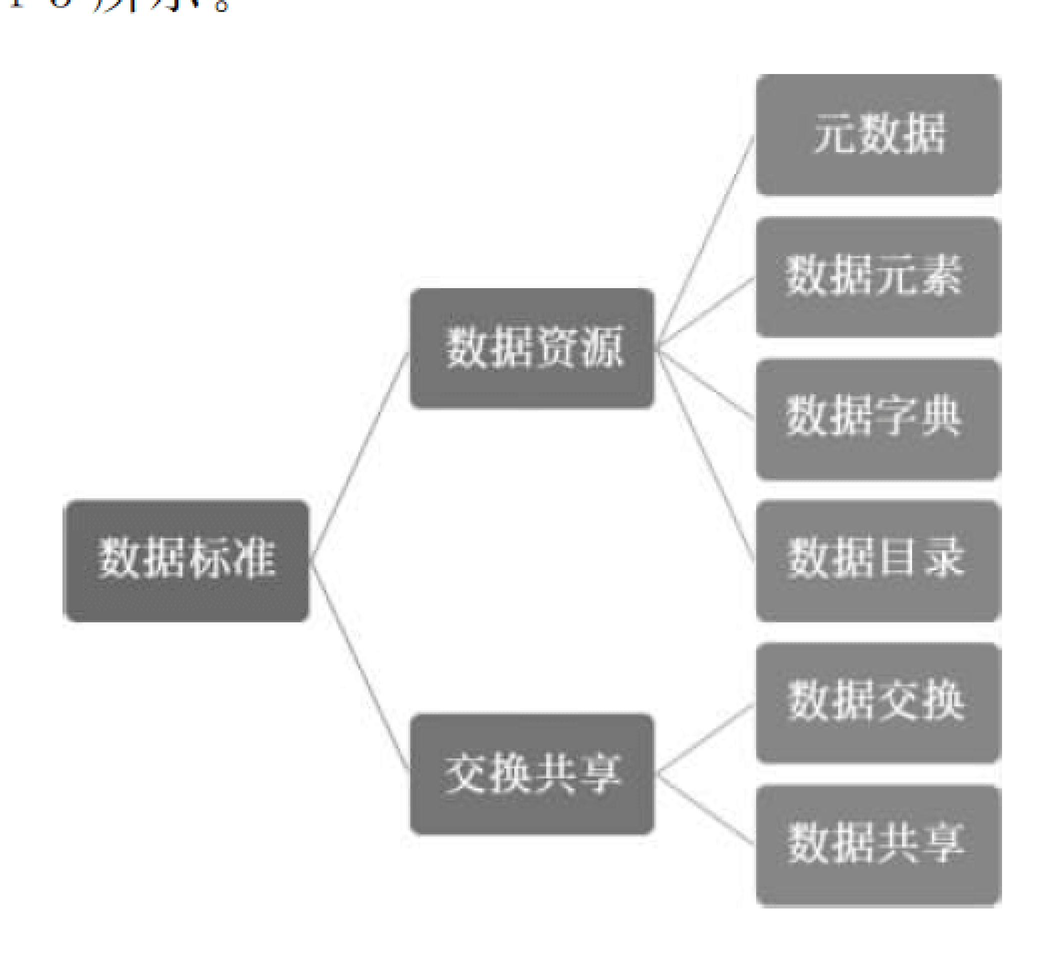


图 4-3　数据标准框架

4.2　数据构建的“知识森林”

大数据成为重要的战略资源，得到国家的高度重视，国务院发布了促进大数据发展行动纲要，提出了发展大数据的指导思想、主要任务和政策机制，“十三五”规划中明确提出十三五期间要提出“要实施大数据战略”，提出“把大数据作为基础性资源，全面实施促进大数据发展行动，加快推进数据资源共享和开发应用，助力产业转型升级和社会创新”，大数据已上升为国家战略。

传统电信行业在 OTT 等互联网产品的冲击下，其话音经营路线已经穷途末路，运营商迫切需要用互联网思维武装自己，才能在这激烈的角逐中不至于被彻底“边缘化”。电信运营商作为数据的生产者，拥有丰富的大数据资源，这些资源优势是其他企业无法企及的，价值挖掘潜力巨大。而拥有如此优质的数据基础，使得运营商在企业、行业、社会等多个层面，都将能够大有作为。中国移动从 2007 年就开始云计算的探索和实践，是国内云计算的先行者和倡导者，积极参与国内外的标准化和产业推动。中国移动通过对数据的挖掘、建模、应用、提升，政府应用，改善转型，在促进企业发展的同时加快推进数据资源共享和开发应用，助力产业转型升级和社会创新。

根据《2015 年全国教育事业发展统计公报》数据，2015 年我国小学生 9926.37 万，初中生 5066.80 万，高中生 4686.61 万，国内中小学生的总人数基数接近两亿，全国中小学数量达到约 32 万所。国家财政性教育经费支出超过两万亿，全国教育信息化年投入超 1000 亿。基于大数据的教学改革逐步开展，若每所学校数据化基础服务费最低三万元/年，则全国 32 万所中小学每年接近一百亿元的基础服务市场规模。全国有两亿的在校学生，每户每月最低 5 元的基础增值业务服务费，则年度个人用户的基础增值服务市场总额大于 120 亿元。

面对迅猛发展的巨大市场，在线教育和移动教育市场既开放又混乱，如今的教育类

---

① 全国信息技术标准化技术委员会大数据标准工作. 大数据标准化白皮书(2016)[R]. 2016.

App五花八门，一个App开发和短期存活的成本并不高，使得教育类App并不愁“量”的问题，学习教育类App数量排名App Store第二位，同时造成了大量低质产品的涌入。通过对市场状况和现有产品的分析，明确了以下需求痛点。

1. 学生（小学、初中、高中）

当前网上教育平台众多，教育资源繁多但良莠不齐，学生需要优质课程的梳理，呈现各学科详细知识结构与脉络（原始知识树）、线上优秀教育资源的评价推荐（科学权威教育资源评价推荐体系）、个性化学习过程的记录与反馈（个性化知识树的构建）、个性化学习计划推荐（自适应学习路径的推荐）、学习成果的反馈（科学的试题库、考核检测体系）、学习成果的个性化呈现（个人知识森林的管理）等。

2. 教师

面对线上众多教育资源，如何为学生选择放心优质的课程、对学习动态的把握、对学生学习成果的掌握、对学习路径进行调整和个性化指导等是教师对线上教育资源的需求。

3. 家长

众多家长流露出既希望孩子进行线上学习又担心的矛盾心理，对孩子线上学习课程不放放心，对线上学习是否有助于学习成绩提升不确定，对孩子线上学习状态难以掌握。

4. 教育资源提供者

现阶段未对教育资源进入线上设立统一的门槛和标准，缺乏对教育资源的评估，致使线上教育资源质量参差不齐，优质教育资源提供方迫切需求安全、高效、信誉优质、用户众多的资源输出平台。

中国移动“和教育”云平台汇聚了北京师范大学、科大讯飞、新东方、好未来、凤凰传媒、华师京城、北京四中等知名教育机构的优质资源，以K12教育为切入点，围绕教师、学生、家长之间真实的客户关系，为客户提供各类教育细分产品，满足不同客户的个性化需求。云平台上线以来，实现汇集全网一千多万条优质教学资源和近三十款精品应用，通过与全国31省对接，在实际建设和运营中发现，当前教育信息化产品对教学过程尚未根本改变，传授知识点不直接精确，且知识点之间缺乏关联，因材施教、个性化学习很难实现。为此开发基于知识点的个性化学习产品“知识森林”。

4.2.1 平台设计

教育领域中的大数据有广义和狭义之分，广义的教育大数据泛指所有来源于日常教育活动中人类的行为数据，具有层级性、时序性和情境性的特征；而狭义的教育大数据是指学习者行为数据，它主要来源于学生管理系统、在线学习平台和课程管理平台等。建立面向中小学学生的和知识森林的平台首先要完成数据的采集和分类。

数据构成如图4-4所示。

数据的来源主要分类线上和线下。线上资源是以中国移动和教育平台教育资源、用户数据为主；线下资源则是对学生、教师、家长、转件、教育企业的访谈数据，也就是对用户需求的调研数据。

按照教育资源和用户行为分层，定量数据和定性数据进行分类，对数据进行脱敏、剔重、聚类等处理。

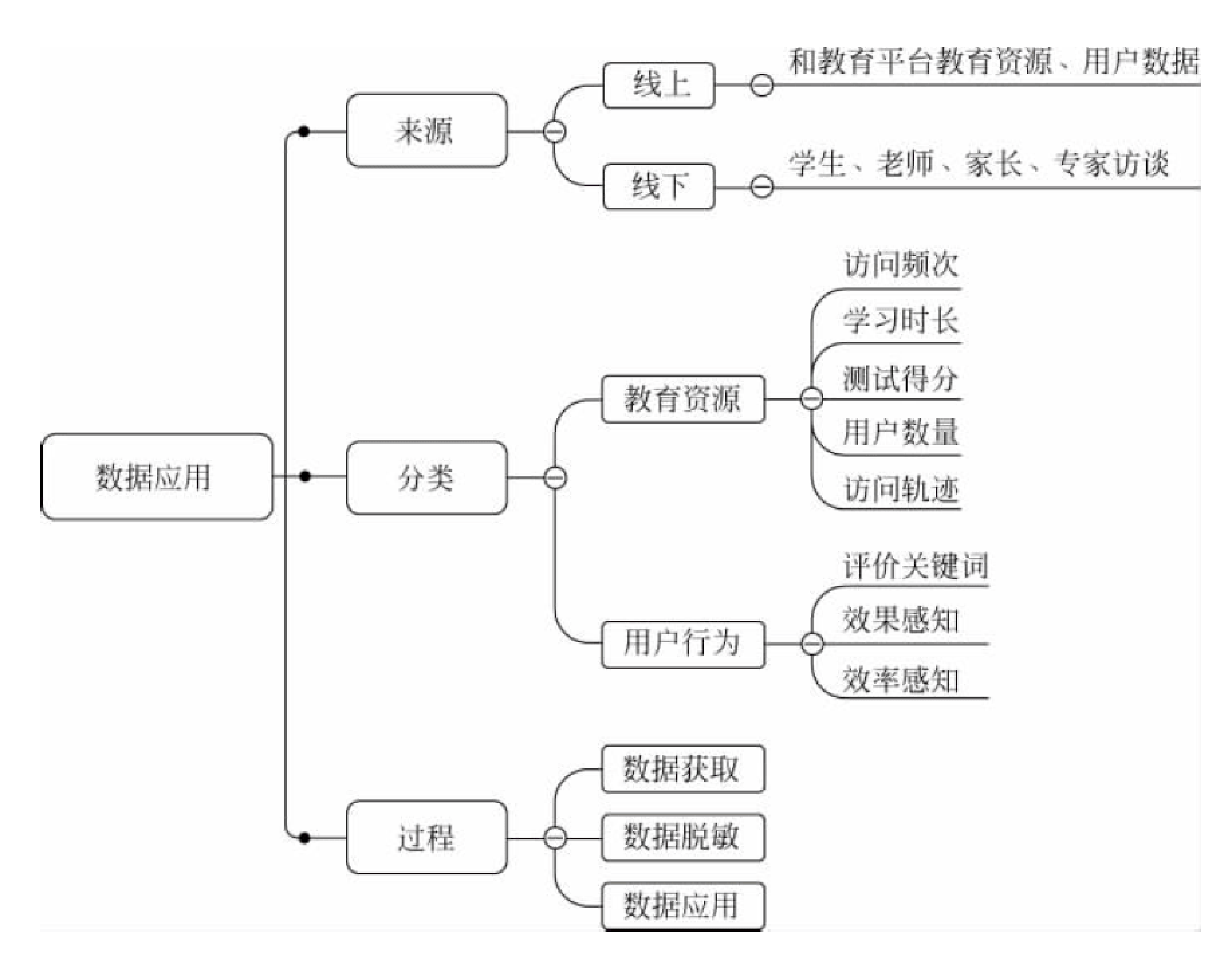


图 4-4　数据构成

数据应用的全过程要按照保证隐私、覆盖全面、实时更新、准确无误的原则进行处理。

4.2.2　实施路径

在“大数据”时代，个人数据得以大量收集、分析，这使得在教育研究中数据研究的价值大幅度提升。同时，教育“大数据”的出现，也使得个性化教学与多样化教学更具可行性。

1. 知识结构为基础构建原始知识树

与和教育平台对接，以系统内教育资源为基础，同时引入重点名校资源并在同一数据标准下实现资源的共享。同时，在对教师教学流程和学生学习反馈进行大量调查的基础上，梳理各学科知识点及考点，构建原始知识树（如图 4-5 所示），实现知识地图形象化呈现。知识点间的关系和脉络可以形象地用树表示，枝干连接与方向表示知识点之间的关系，每片树叶代表不同的知识点。建立各学科原始知识树，每一学科一棵树，每一知识点一片叶，进行分区，主科为乔木区、副科为果树区、兴趣为花丛灌木区，未学习者单击学习为枯树枯叶，且以种子形态展示，但单击时能以树木生长的过程显示知识关联和潜在学习过程，如图 4-6 所示。

基于已有的学科知识本体，允许通过半自动化和人工编辑的方式逐步建立学科知识之间的语义关联关系，并采用自上而下以及自下而上两种关联进化的思路，实现学科知识的语义关联及其进化。自上而下是指由学科专家对平台中新增的学科知识关系进行人工审核以及手动增加学科知识之间的关系，由学科专家审核或新增的学科知识关系直接被系统采纳；而自下而上是指由普通用户根据现实需要对平台中尚未定义的学科知识或学科知识之间的关系，以及尚未标注的学科知识之间的关系进行添加，或是基于平台推理引擎所产生的相关学科知识和学科知识关系等。既包括已建立关联关系的学习知识点，也包括待建立关联的

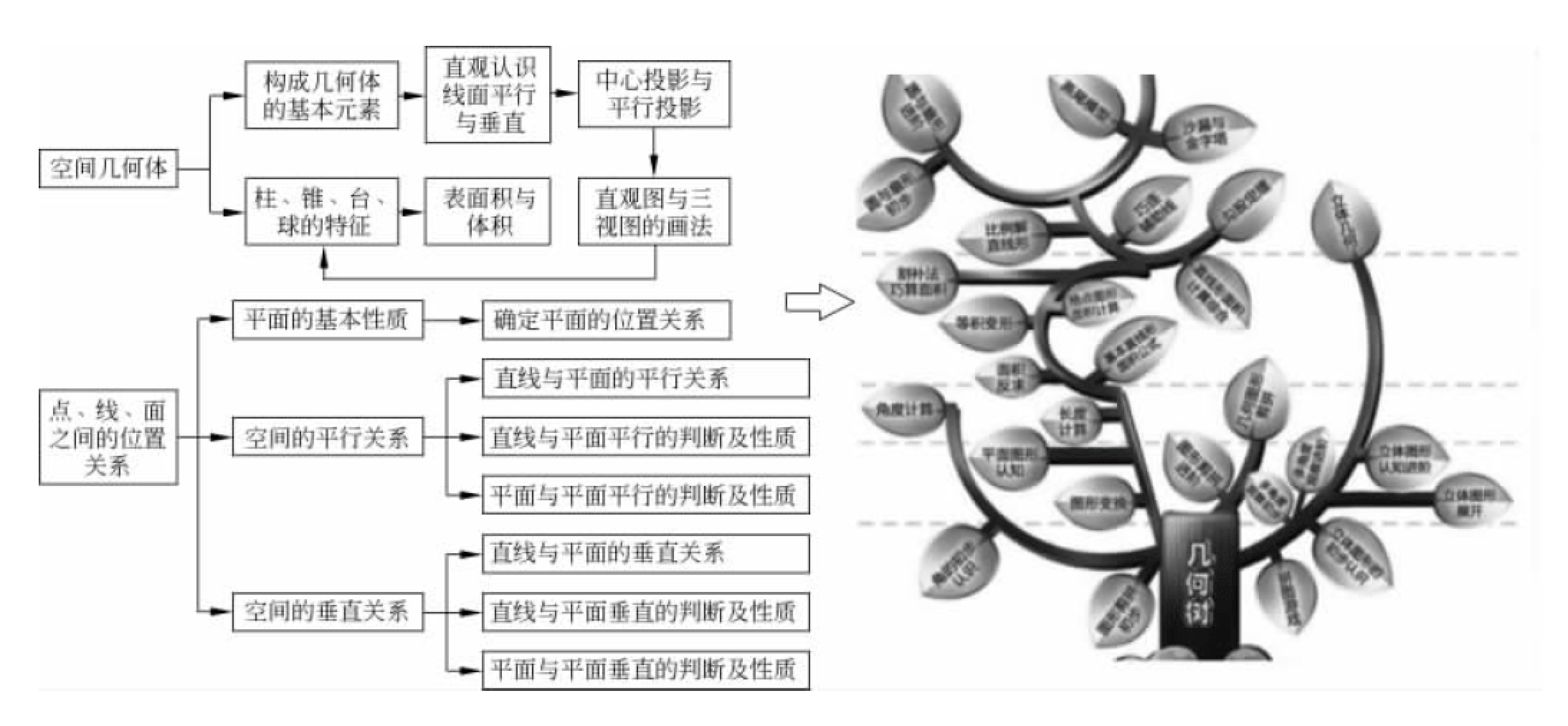


图 4-5　知识树

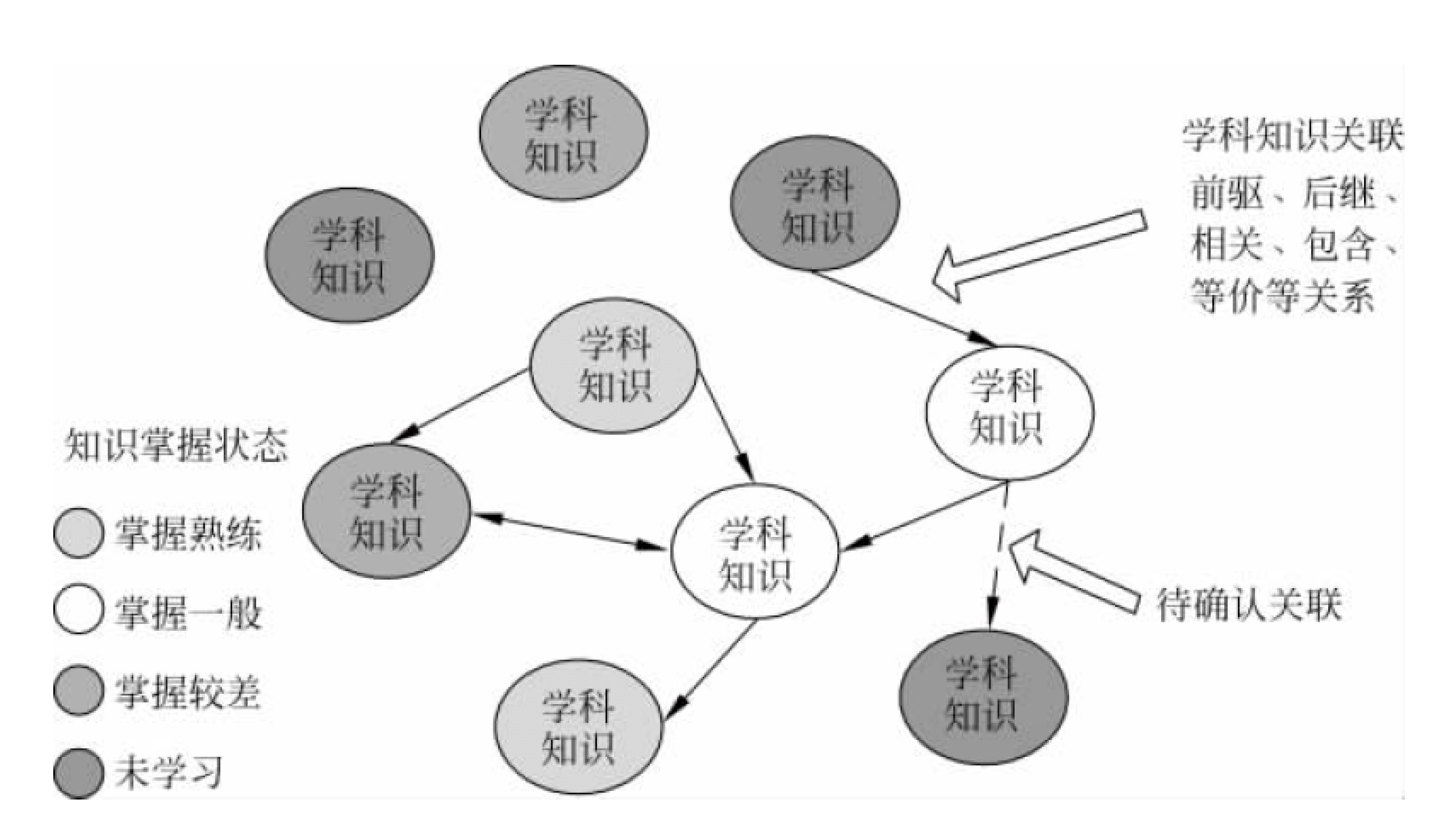


图 4-6　知识网络

学习知识点，同时对于课程知识群中的学习知识点及其之间已建立的关联关系还能够进行持续更新和进化。利用基于学科知识本体所建立的学科知识关联，形成包含学科知识及其之间相互关联关系的原始知识树。

2. 需求感知，预判用户个性化需求，实现自动化学习

需求感知过程如图 4-7 所示。

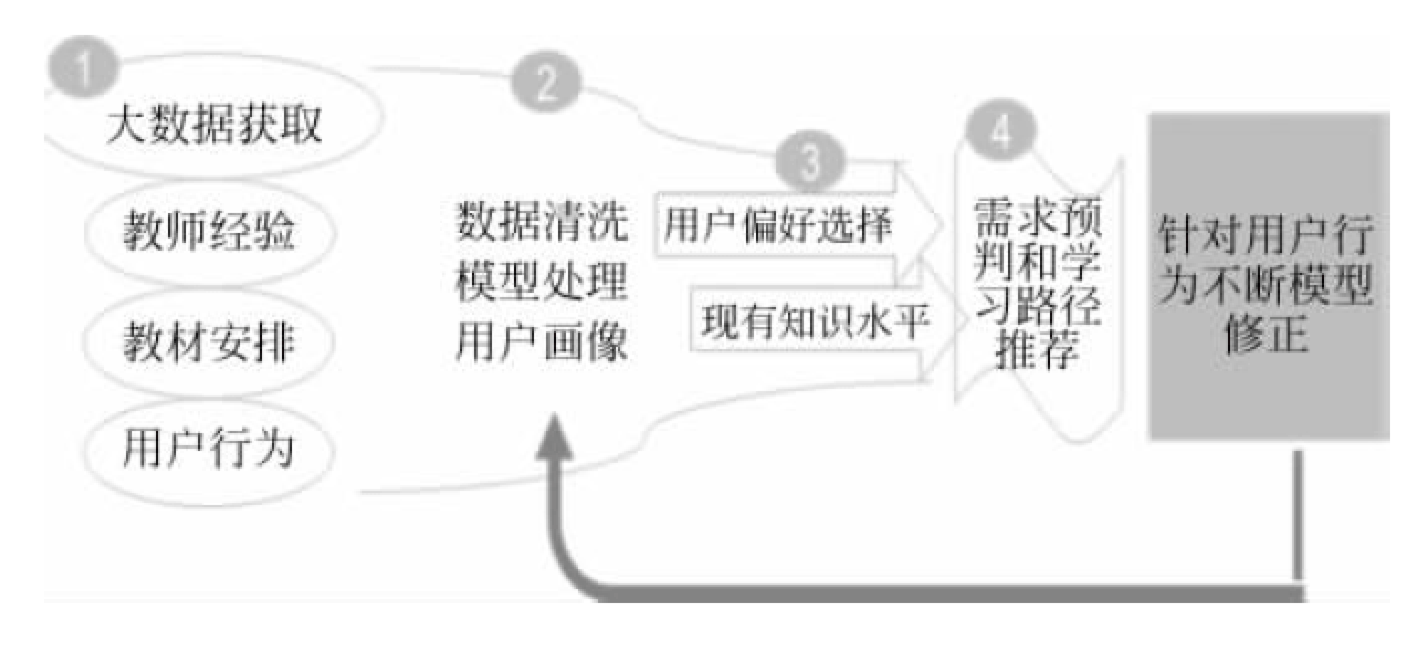


图 4-7　需求感知

（1）通过对老师经验数据、教科书设计路径、线上用户以往行为数据的数据清洗、模型处理、用户画像建立用户行为数据库和分析平台，提供用户需求预判和学习路径推荐，同时对知识点关联关系进行提示。

（2）根据学生的实际表现，建模，将大数据平移到新用户，根据用户的实际反馈来不断小步快跑迭代修正模型。

（3）通过学生自主晒出经历，经验分享，构建用户知识需求感知的互动氛围。

通过需求感知功能，预判用户个性化需求，进行学习路径推荐，以树叶摇曳形式呈现，同时对知识点关联关系进行提示，可爱的树叶像招手一样吸引用户学习。

3. 通过用户行为数据实现个性化知识树构建

如图 4-8 所示为一棵个性化知识树。

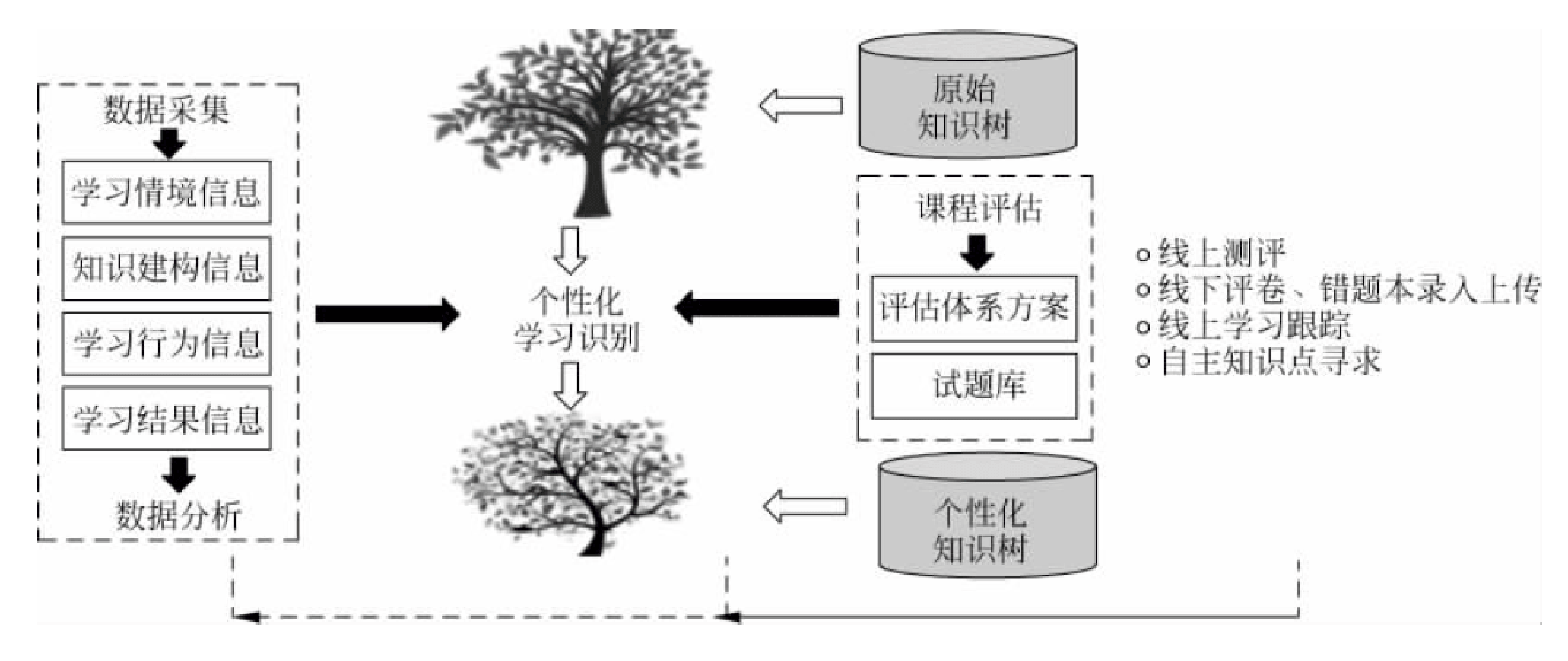


图 4-8　个性化知识树

传统的在线学习过程数据采集主要通过监控和跟踪学习者的数据库访问记录或 Web 日志文件来实现。综合对现有在线学习行为数据采集研究的分析，可以发现目前主要存在以下两方面的不足：一是忽视学习者对学习内容本身贡献的行为数据采集；二是学习行为数据采集与分析模型很少关注学习者当前的学习情境信息，如地理位置、气候、学习终端、网络环境等。

通过对学习者行为数据的跟踪和分析，进行个性化学习识别，从而实现个性化知识树的构建。

学习情境信息：学习者初始能力水平信息、学习终端设备信息、地理位置信息以及时间信息等影响学习者获取和运用知识的一切要素。

知识建构信息：学习内容与学习资源的编辑、审核、批注、分享和传播等对学习内容和学习资源进行再创造的贡献数据。

学习行为信息：以学习活动为核心，主要包括浏览学习内容与资源、参与学习活动、完成学习任务等过程性的行为数据。

学习结果信息：主要包括学习时长、完成活动质量、测试成绩等成果性数据。

4. 个性化知识树再学习，教师、家长动态把握及有效干预

如图 4-9 所示，个性化知识树建构过程中掌握程度不同知识点树叶呈现不同颜色，并以

树叶摇曳形式再次推荐优化学习路径；优化学习后再次提供效果测评，根据测评结果，树叶颜色发生变化。教师（家长）通过姓名查看详情可以清楚查阅某人的学习动态，包括学习时长、学习内容及路径、学习成果评估等。教师（家长）根据学生知识点掌握情况，发现知识结构薄弱方向，及时对学习路径进行调整和个性化指导，可点对点向学生发送学习任务安排、学习提醒及学习跟踪等。

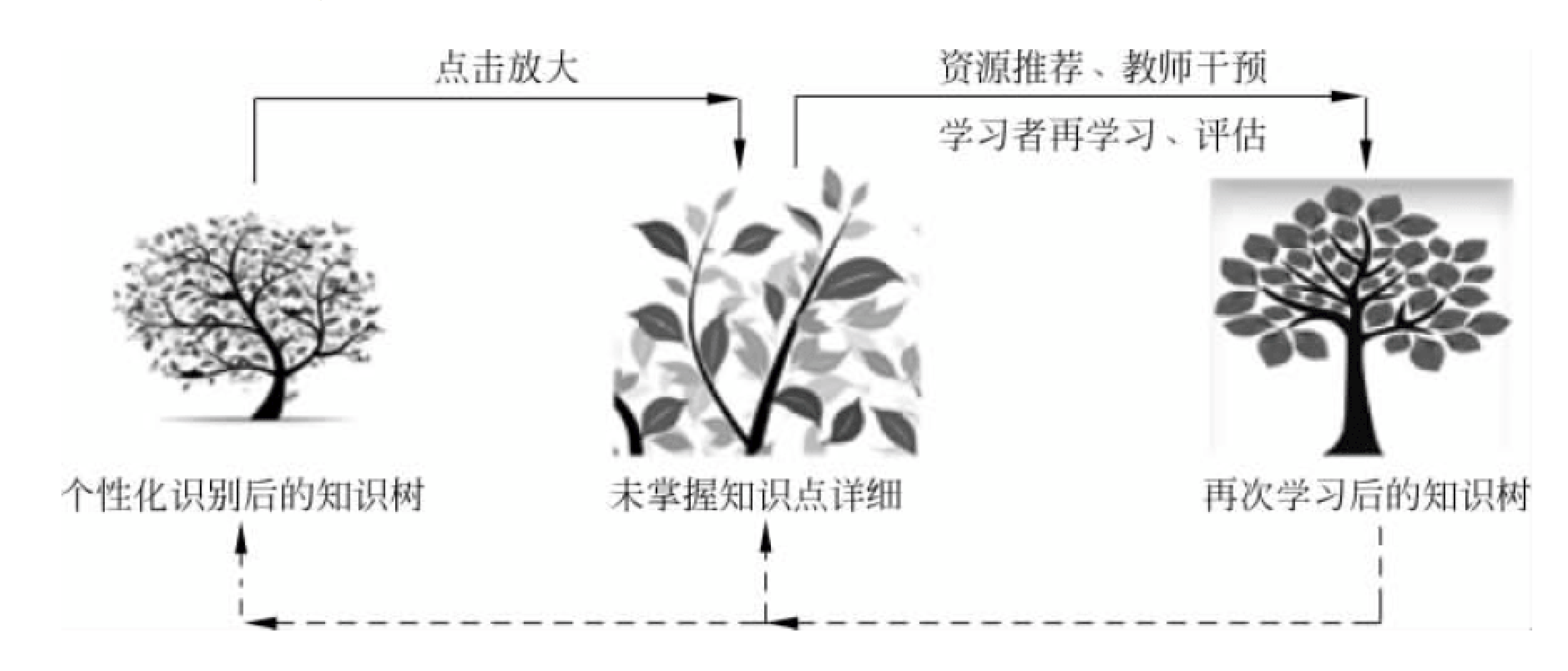


图 4-9　再学习时的知识树“成长”

5. 线上教育资源评估

当前线上教育资源的使用效率和效果，并未随着资源数量而同步提高。资源的重复建设、数量庞杂、良莠不齐、重量轻质，是突出的现象。数量庞大杂乱，使得学习者无从选择；质量良莠不齐，使得学习者无从辨识；重形式轻理念，脱离传统模式，使得学习模式不可能发生根本的变化。面对庞杂的、未经甄别、未有指导建议的海量资源，学习者只会更加困惑、迷惘，无从选择或盲目选择，严重影响着学习者的学习效率和学习质量，影响着网络教育的持续、良性发展。

如图 4-10 所示，设定专家评估子体系和学习者评估子体系分别从专家的学术角度和学习者的使用角度对数字资源进行评估。专家评估子体系包括专家标准化意见和主评专家意见。标准化意见包括若干个评估维度、评估项以及权重；主评专家意见则是相对个性化的特色点评。学习者评估子体系包括主观评估和客观评估。主观评估指学习者对于具体数字资源的主观点评；客观评估则指基于学习者网络学习的行为习惯等数据挖掘。对教育资源进行排序推荐，减少学生学习盲目性。随着数据积累和学生反馈可将该模型用到全部线上教育资源。

4.2.3　案例分析

1. 历史数据及用户行为数据应用

充分利用“和教育”现有教育资源和用户资源大数据，初步实现了系统自动挖掘，根据系统内现有课程资源构建原始知识书，基于平台推理引擎所产生的相关学科知识和学科知识关系构建原始知识树；通过教师、教材、用户行为数据实现了需求预判和个性化学习，通过不断迭代提升了精度。

2. 个性化学习与指导

学生可按照系统推荐进行学习，教师和家长可以及时了解学生对知识点的掌握情况，发

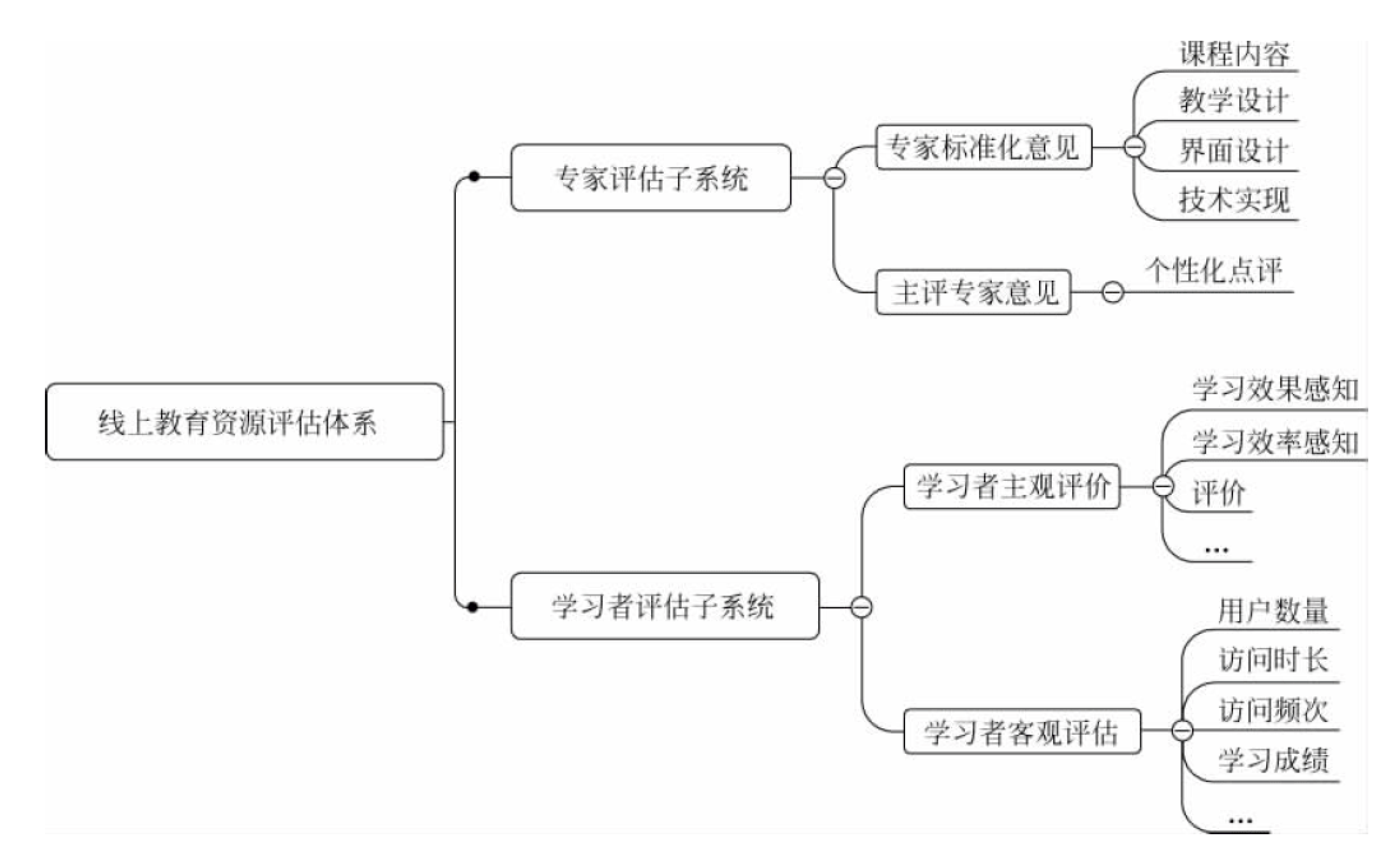


图4-10　线上评估

现薄弱方向，进行个性化指导。学生按图索骥快速掌握课程涉及知识点，形成学生个性化学习路径和知识图谱，实现个性化学习，解决了当前学生、家长及教师的痛点需求。

3. 线上教育资源评估

现阶段未对教育资源进入线上设立统一的门槛和标准，缺乏对教育资源的评估，致使线上教育资源质量参差不齐。评估体系使得资源相关方能够便捷地看到权威专家的定性、定量评定，直观地看到同类网络资源排行榜，也能够看到学习者的主观评价以及客观评价，掌握某门具体网络课程的学习者数量和受欢迎程度，掌握某门课程的具体特色以及网络学习的时间周期等规律，从而为资源使用者提供指导，为资源制作者提供建议，为资源运营者提供依据，为资源研究者提供参考，对于引导学习者、资源制作者、资源运营者以及资源研究者有着重要意义，对于网络教育的发展将起到积极的推进作用。

4. 中国移动大数据应用现存问题

由于技术、数据系统限制，用户隐私和商业模式不明确等问题，目前大数据应用只处在探索阶段，主要遇到以下问题。

(1) 系统分散建设，难以实现资源共享。经营分析、信令监测、上网日志留存等众多数据系统分专业建设，其中部分系统还分省建设，造成资源无法共享。

(2) 数据处理种类多，单一技术难以实现。各大数据系统数据模型不统一，只具备结构化数据处理能力，无法支持非结构化、半结构化数据处理，无法满足互联网业务发展要求。

(3) 如何避免隐私泄露。人们对于隐私问题越来越重视，数据公司掌握大量数据和数据制造者要求隐私权之间的矛盾，使得大数据应用变得困难。

(4) 尚未确立商业运营模式。掌握的数据很多，但是这些数据应该怎样应用、给谁用、应用收益是否可以抵消数据开发分析的成本？

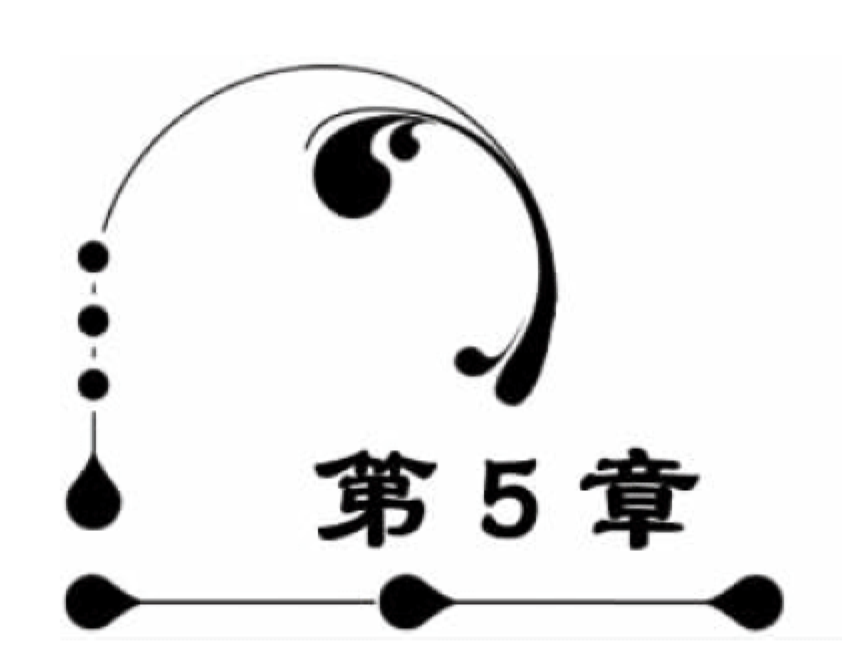
5. 中国移动大数据应用发展方向

第一，在市场层面，通过大数据分析用户行为，改进产品设计，并通过用户偏好分析，及时、准确且有针对性地开展营销与维系，不断改善用户体验，增加用户信息消费以及对运营商的黏度；第二，在网络层面，通过大数据分析网络流量、流向变化趋势，及时调整资源配置，同时还可以分析网络日志，进行全网优化，不断提升网络质量和网络利用率；第三，在企业经营层面，可以通过业务、资源、财务等各类数据的综合分析，快速准确地确定公司经营管理和市场竞争策略；第四，在业务创新层面，在保障用户隐私的前提下，可以对数据进行深度加工，对外提供数据分析服务，为企业创造新的价值。

第三篇　大数据，价值创新的土壤

亚洲首富孙正义说过：“每个人都有大脑，但不是每个人都有智慧；每个人都有眼睛，但不是每个人都有眼光；每个人都有双手，但不是每双手都能把握机会；机会永远都是给那些有智慧、有眼光、有勇气、有准备的成功者而准备的！”大数据的应用可以成为每个人的最强大脑，数据中所蕴含的“智慧”和可预见的“眼光”就是为有勇气的人做准备最好的机会。

“数据不是黄金，不是石油，却是未来新经济发展的土壤。在这个前所未有的时代，大数据成为一种重要资源，推动创业创新。”——涂子沛



大数据精准营销

随着互联网的日益普及，人们对互联网技术的利用率越来越高，由此而来的大数据对社会的各行各业都带来很大变化，人们正步入大数据时代。在企业营销中，如何利用大数据发掘用户需求、精准找到目标用户群从而形成强有力的营销方案，是为其带来发展机遇的关键所在。

5.1 大数据营销

大数据营销是指通过互联网采集大量的行为数据，首先帮助广告主找出目标受众，以此对广告投放的内容、时间、形式等进行预判与调配，并最终完成广告投放的营销过程。

大数据营销，随着数字生活空间的普及，全球的信息总量正呈现爆炸式增长。基于这个趋势之上的，是大数据、云计算等新概念和新范式的广泛兴起，它们无疑正引领着新一轮的互联网风潮。

5.1.1 精准营销

1999 年，美国的莱斯特・伟门提出了精准营销（Precision Marketing）的概念。2005 年，菲利普・科特勒（Philip Kotler）在其全球巡回演讲论坛上宣布，精准营销将是营销传播的新趋势。科特勒在其畅销书 *Principles of Marketing* 中，首次将基于互联网的精准营销理论融入其中，他认为日新月异的科技，使一些公司勇于从传统的大众传媒沟通方式转移到更加有针对性目标市场的互动模式，以此来不断提高沟通的效果和效率。并提出"对于营销来说，将沟通个性化，在正确的时间，对正确的人，表达并且做出正确的事情，是至关重要的。"

简单来说，精准营销就是要做到 5 个合适：在合适的时间、合适的地点，将合适的产品以合适的方式提供给合适的人，如图 5-1 所示。这与人际交往中的男女恋爱是比较相似的，必须是在对的时间遇到对的人。

1. 精准营销的特点

（1）精准营销真正贯彻了消费者导向的基本原则。4C 理论的核心思想，便是企业的全部行为都要以消费者需求和欲望为基本导向。精准营销作为这一大背景下的产物，强调的仍然是比竞争对手更及时、更有效地了解并传递目标市场所期待的满足。这样，企业要迅速而准确地掌握市场需求，则离消费者越近越好。这是由于，一方面，信息经过多个环节的传

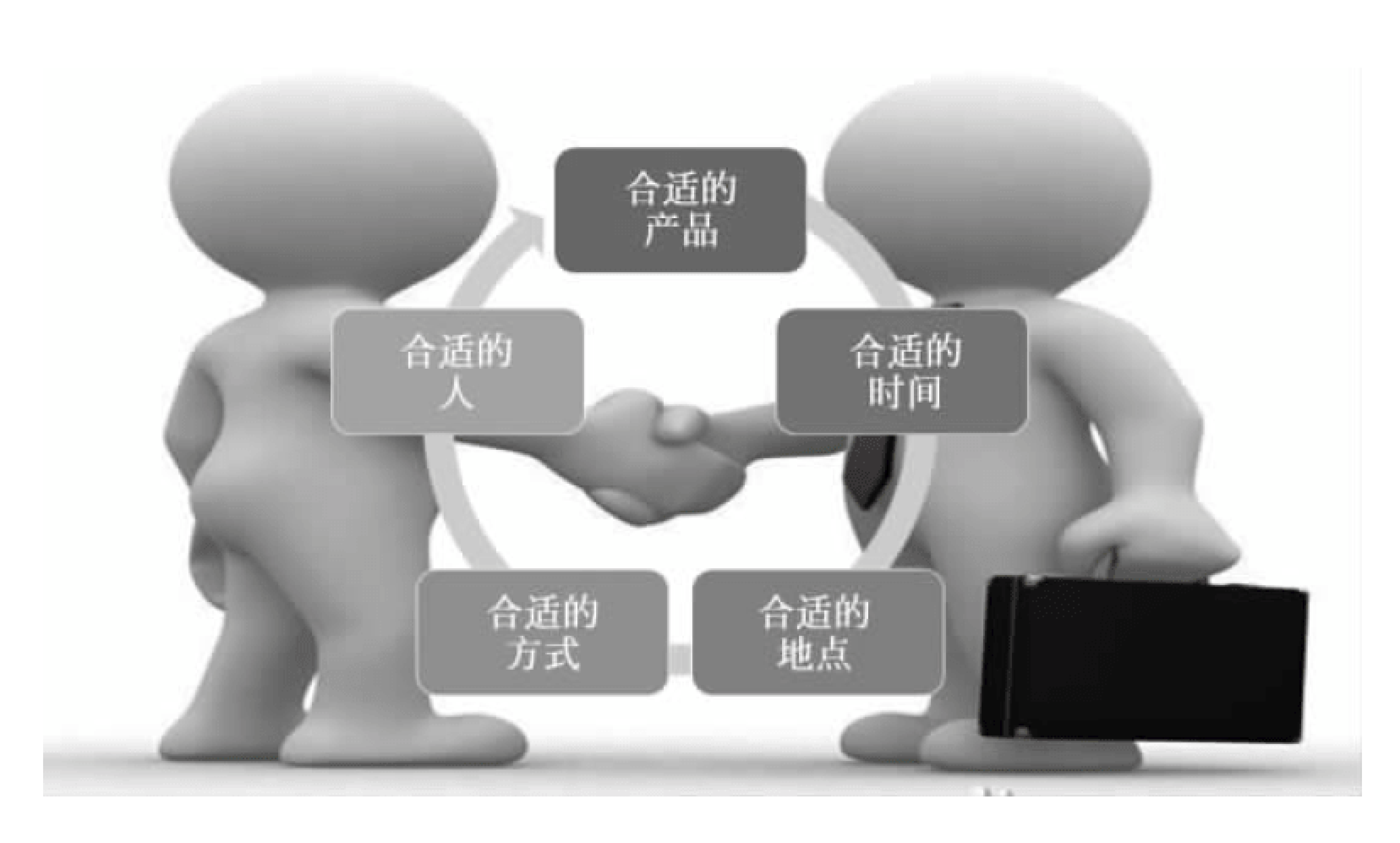


图 5-1　精准营销

播、过滤，必然带来自然失真，这是由知觉的选择性注意、选择性理解、选择性记忆、选择性反馈和选择性接受所决定的；另一方面，由于各环节主体利益的不同，他们往往出于自身利益的需要而过分夸大或缩小信息，从而带来信息的人为失真。精准营销绕过复杂的中间环节，直接面对消费者，通过各种现代化信息传播工具与消费者进行直接沟通，从而避免了信息的失真，可以比较准确地了解和掌握他们的需求和欲望。

（2）精准营销降低了消费者的满足成本。精准营销是渠道最短的一种营销方式，由于减少了流转环节，节省了昂贵的店铺租金，使营销成本大为降低，又由于其完善的订货配送服务系统，使购买的其他成本也相应减少，因而降低了满足成本。

（3）精准营销方便了顾客购买。精准营销商经常向顾客提供大量的商品和服务信息，顾客不出家门就能购得所需物品，减少了顾客购物的麻烦，增进了购物的便利性。精准营销实现了与顾客的双向互动沟通，这是精准营销与传统营销最明显的区别之一。

2. 精准营销的实现

大数据时代之前，企业一般仅能从 CRM（Customer Relationship Management，客户关系管理）或 BI（Business Intelligence，商务智能）系统中获得顾客信息、市场促销、广告活动、展览等结构化数据以及企业官网的一些数据。但这些信息只能达到企业正常营销管理需求的 10%，还需要其他 85% 的数据，诸如社交媒体数据、邮件数据、地理位置、音视频等这类以图片、视频等方式存在的信息数据等，才足够给出一个重要洞察和发现规律。大数据技术进一步提高了算法和机器分析的作用，使得这类数据在竞争激烈的市场中日显宝贵、作用突出。

图 5-2 所示，大数据时代，实现精准营销主要有如下三部曲。

第一步：知己，意味着知道自己产品的定位是什么，产品的卖点是什么等。

第二步：知彼，简单地说就是清楚竞争对手的情况，清楚目标用户的情况。

第三步：作战，对不同的对象采取不同的策略，直击痛点，实现转化。

1）精准的市场定位

市场营销中有两个著名理论：一个是 2∶8 法则，即企业 80% 的收益来自 20% 的用户，不同的客户会给企业带来不同的价值；另一个是“长尾理论”，只要存储和流通的渠道足够

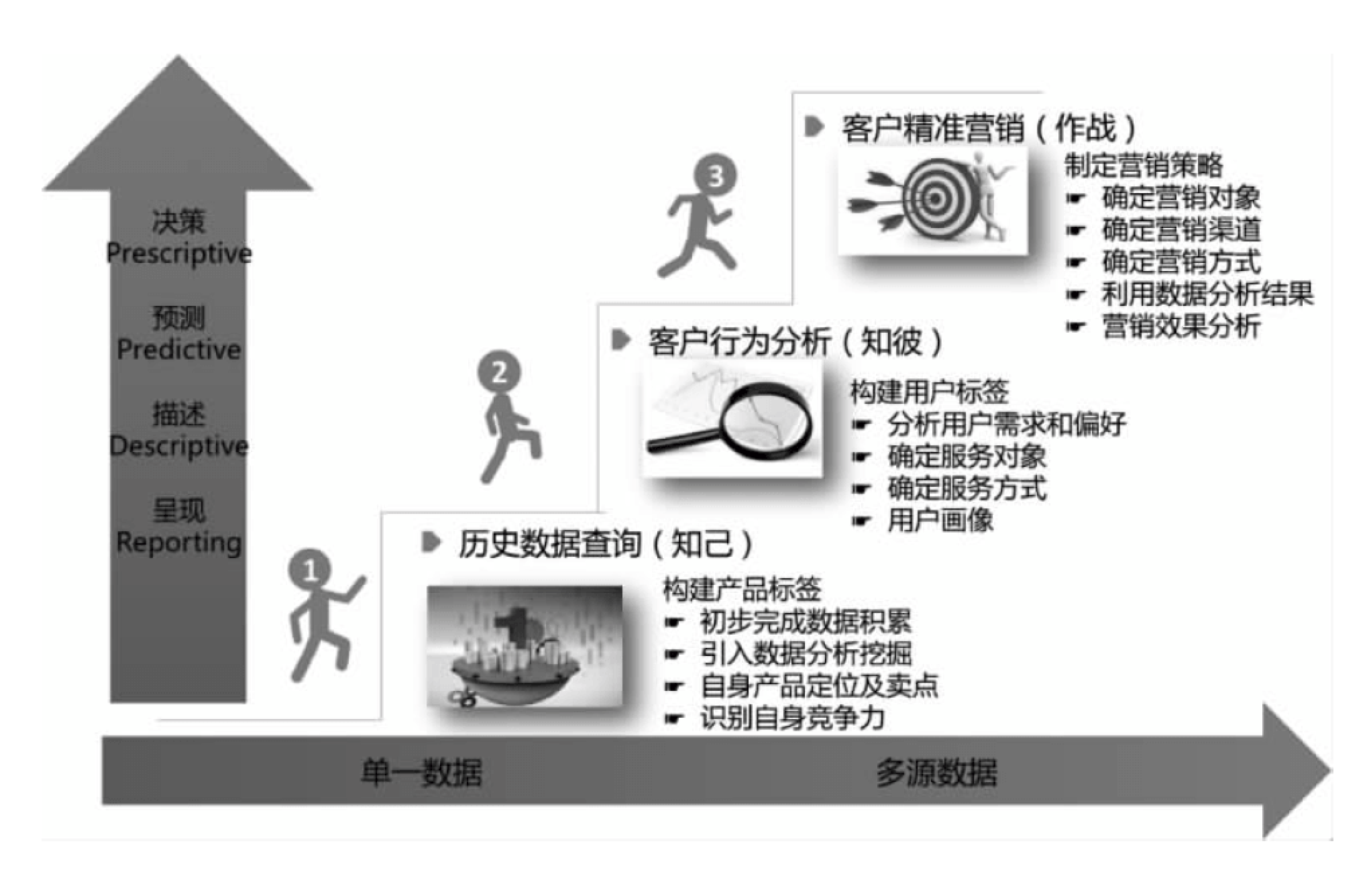


图 5-2　精准营销的实现步骤

大，那些之前被认为冷门或不易销售的产品共同占据的市场份额就可以和那些数量不多的热卖品所占据的市场份额相匹敌甚至更大。

当企业准备将产品推向市场时，必须先找到准确的市场定位，我的产品是什么？它的客户到底是哪些人？如何能够精确地找到目标客户？这些都是精准的市场定位所必须思考的。

2）精准的客户分析

数据的价值在于分析，利用大数据技术，可以对互联网中的用户行为，即用户的网络消费习惯和行为模式进行深入研究。

一是互联网站利用 Cookie 技术捕捉和定位用户 ID，同时锁定该 ID，追踪他在其他类型的网站的行为轨迹，将零散片段拼合出该用户的特征，再根据用户的注册身份和互动分享内容，判断其身份特征、生活方式和关系圈子，最后再借助移动互联网技术分析其实时的地理位置绘制出更立体、更实时的用户画像。

二是建立企业与用户间的新型互动关系，打破以往的“自上而下”“一对多”的线性关系，建立个体间的“一对一”实时互动。①

3）精准的策略与更高的用户体验

大数据根据消费者的“行为轨迹”，分析其消费需求，能够进一步判断其关联需求，挖掘其潜在需求，对其消费需求进行预测；再通过具有针对性的关联推荐，促成有效购买和消费。零售业巨头沃尔玛通过大量消费者购买记录分析，发现男性顾客在购买婴儿尿布时，常常会顺便搭配几瓶啤酒来犒劳自己，于是推出“啤酒和尿布”捆绑销售的促销手段，直接带动这两样商品的销量，成为大数据营销的经典案例。

在以市场导向、消费者为中心的营销新时代，要想获得收益，企业就必须关注客户价值。

① 胡江涛. 大数据营销：从精准到实效[J]. 鄖阳师范高等专科学校学报，2014，34(6)：56-59.

客户价值的实现才可能带给企业丰厚的利润和回报。当然，只有当客户的需求转化为公司价值时，企业才是真正满足了客户需求，而这必须通过客户体验，来表明他的需求。由此可见，以消费为导向、关注消费个体体验就是精准营销中要实现更高的客户体验的真谛。

5.1.2 精准广告

广告，即广而告之之意。广告是为了某种特定的需要，通过一定形式的媒体，公开而广泛地向公众传递信息的宣传手段。宽通广告全国运营总监宋琼表示，我们生活一个广告充斥的年代，在众多的屏幕当中，电视媒体能够覆盖到中国大多数的居民，在白天和晚上所表现出的数据是相类似的，互联网改变了长期以来仅作为传统电视补充的角色定位，拓展了自身的商业价值。如何科学、有效地找到目标受众，如何合理地整合媒介资源，如何真正地精准投放广告是产业链里面各个环节思想的关键问题。

2006 年，百度风向标般洞悉了网络搜索给予互联网界乃至全球的重大影响，在首届百度世界大会上公布了创新的广告形式——精准广告，以让广告呈现且仅呈现在想要呈现的人面前为目标，如图 5-3 所示。

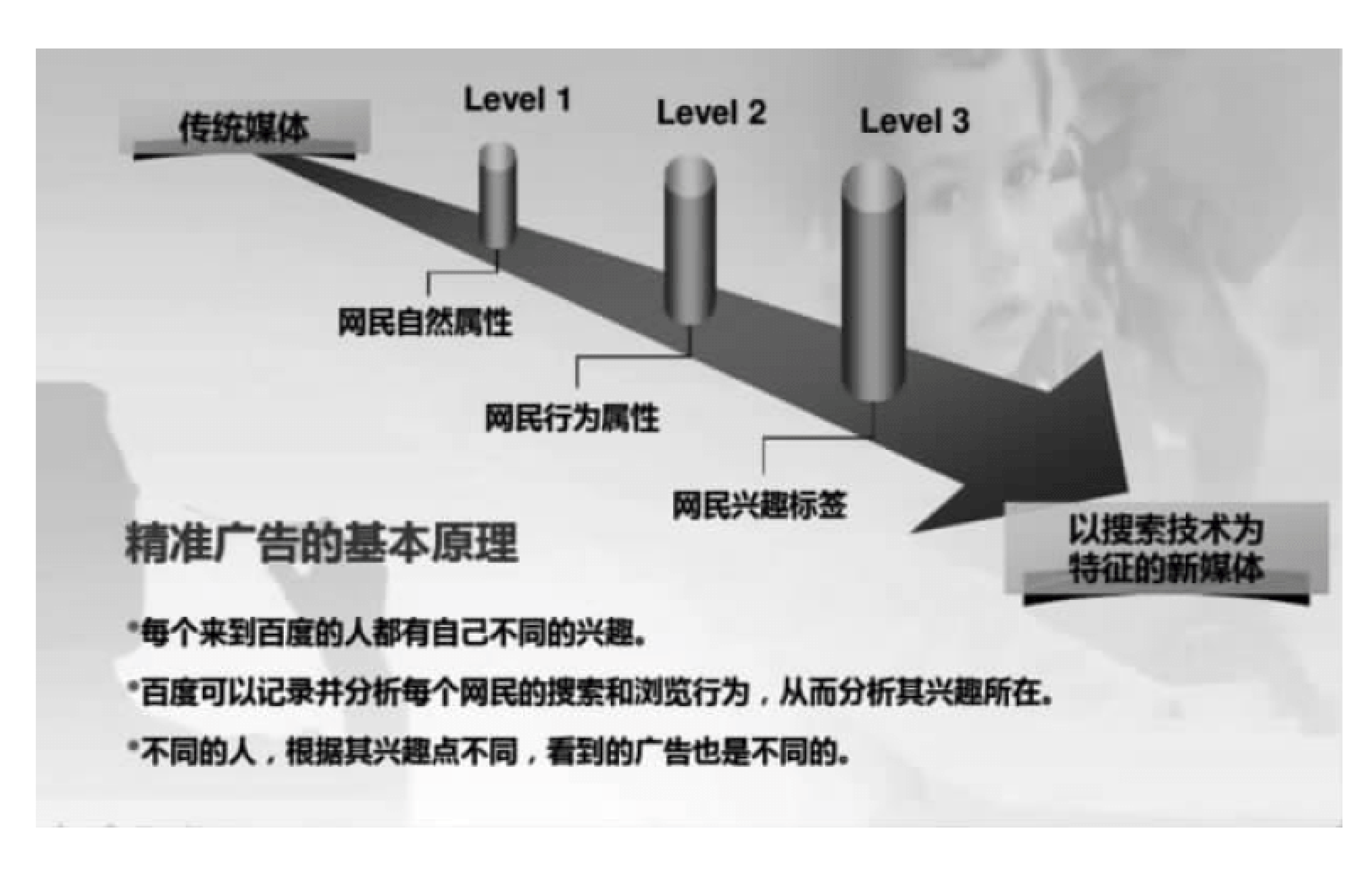


图 5-3 百度精准广告的实现原理

大数据与云计算时代的到来，为海量数据存储、处理提供了强大的技术驱动与支撑。通过对浏览器 Cookie（Cookie 就是网上服务器为了辨认某用户的计算机，暂时存放在该用户计算机上的一点儿资料）、用户注册数据、用户行为数据等记录的抓取，分析用户的消费者属性数据（包括基础属性、行为属性和心理属性），再利用其他途径丰富用户数据维度，定量与定性结合，细致分析每一位用户的基础信息、行为与心理特征，精准广告应用的一系列技术可以实现广告的个性化定制投放。

5.2 实时竞价广告

随着互联网的快速发展，横幅、弹窗、悬浮窗、文本、图片、视频等各种类型广告扑面而来，广告行业也越来越呈现出媒体多元化、用户碎片化等特点。对广告精准化的要求一直以来都是广告业最关注的核心问题，即如何将互联网广告在合适的时间以有效的方式传递给

目标人群。

中国核心企业网络广告投放市场细分如图 5-4 所示。

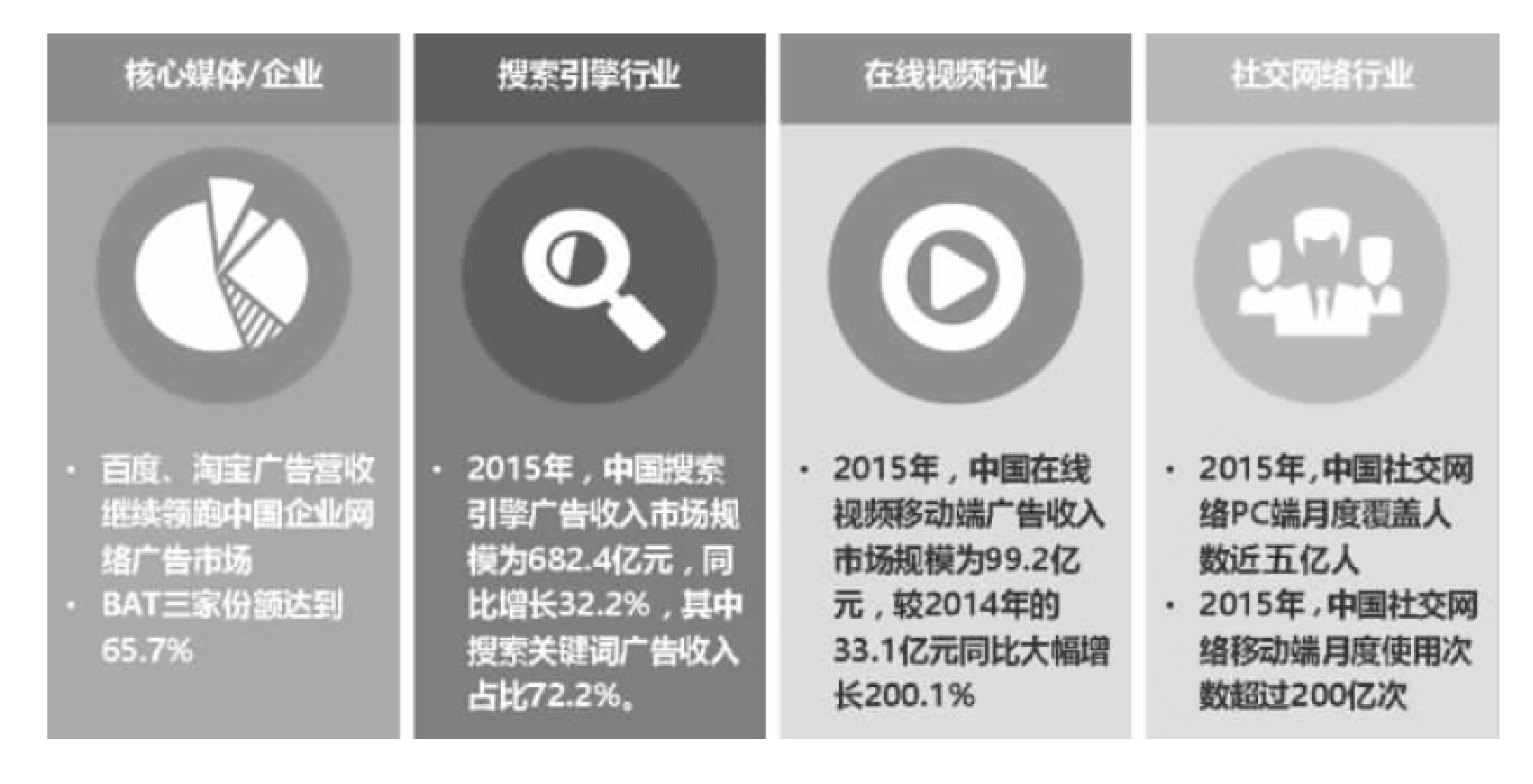


图 5-4　中国核心企业网络广告投放市场细分

广告主投放广告时，着重考虑成本收益的问题就凸显出来了，为了在适当的情况下以合理的竞价策略和模式来投放广告，以保证投入的收益，便催生了精准营销下的 RTB 广告竞价模式。

RTB(Real Time Bidding)是一种以互联网大数据为背景的实时网络广告竞价购买模式。它融合了大数据技术，将数据挖掘和预测应用到展示广告上，与传统购买形式相比，它是在每个广告展示曝光的基础上进行竞价，即对每一个页面访问(Page View，PV)进行竞价，谁出价高，谁的广告就会被这个 PV 看到。

RTB 的兴起和发展有效地降低了广告投放成本，提高了媒体的收益率和广告主的投资回报率，避免了无效到达，从粗放的购买固定广告位全面曝光转换成面向具体独立用户的精准投放。例如，单独访客记录(Cookie)、IP 地址(Internet Protocol Address)或用户身份(ID)，在一定程度上实现了互联网广告用户的个性化投放管理。

5.2.1　RTB 广告投放关键技术

RTB 实时竞价是一种利用人群定向，在数以百万计的网站上，针对每一个用户的网上行为进行评估以及出价的竞价技术。这种实时竞价模式允许广告卖价根据活动目标、目标人群以及费用门槛等因素，对每一个广告及每次广告展示的费用进行竞价。当用户每次打开网页时，广告主根据广告交易平台提供的用户数据进行分析并判断，随后决定是否对该次展现竞价，以及出多少价钱去竞争，如果竞价成功，就将会在此访问者的访问过程中出现该广告主所投放的广告。在这个过程中，广告主本身对数据的理解、使用和分析存在一定的局限性和制约，还需要一定的第三方平台支持。

1. Web 挖掘

Web(网络)挖掘是指应用数据挖掘技术，针对互联网页面数据进行分析与处理，提取出隐含在其中的、人们事先不知道的但是潜在有用的信息和知识的过程。Web 挖掘分为三类：内容挖掘、结构挖掘和使用挖掘。在 RTB 广告投放模式中，主要应用内容挖掘来搜索和抓取关键字和页面重要信息，以此来充当沟通用户与广告主的一个桥梁，完成广告的推荐

排名。

Web 数据纷繁复杂，首先需要对数据进行清洗和预处理，主要包含用户识别、会话识别和路径补充。用户一般是通过 Cookie 来识别的，不同 IP、不同浏览器都被认为是不同的用户，在能够得到用户的唯一专属数据的情况下，也能够跨终端、IP 等进行用户的识别。

2. 协同过滤

协同过滤是基于集体智慧的一种典型的方法，通过与个体相似的其他多个用户的兴趣和爱好来推测目标用户的喜好，通过群体合作的方法来帮助别人找到所需的内容。协同过滤常被用于电子商务推荐系统中，分析用户兴趣，在用户群中找到指定用户的相似（兴趣）用户，综合这些相似用户对某一信息的评价，形成系统针对该指定用户对于此信息的喜好程度预测。

在分析用户行为和预测用户喜好时，主要通过搜索关键字和他所浏览页面中包含的关键字提取出用户行为偏好，结合 Web 页数据和用户历史数据来构建用户画像，综合用户画像和行为偏好，进行广告项目的匹配，并充分考虑用户、兴趣、广告三个方面的因素。协同过滤以其出色的速度和健壮性，在全球互联网领域炙手可热。

5.2.2 RTB 的生态圈

RTB 的产生，标志着广告行业从卖广告位、卖时间到卖专属用户、卖场景的转变，从固定价格（如按展示量 CPM、按点击量 CPC 付费等模式）到实时竞价的智能化投放，从单一系统到将受众、媒体、广告主三方整合的竞价投放模式，满足了企业广告主个性化、精确化、多样化的需求，同时也考虑用户需求，注重用户体验，实现了品牌与效果的统一发展。

虽然 RTB 可以定位人群，但也需要像搜索关键字一样定位用户意图，所以其效果的提升也只是相对于一般展示广告而言。RTB 的核心优势主要有两点：实时竞价；智能投放。这都需要有相应的平台技术，而国内对于 RTB 算法发展尚处于一个相对不太完善的阶段。在广告的投放中，广告竞价并不是一次就结束了，而是一个持续优化的过程，其中对于竞价算法和推荐优化也都有更高的要求。

RTB 广告的出现再次证实了广告是“科学的”，RTB 得以迅速发展，很大程度上是凭借“大数据”的支持。抛开广告内容设计因素，这里主要对广告的中下游阶段——广告媒介购买及投放过程进行分析。通过 Cookie、IP、ID、点击行为等数据，借助云计算等技术可以对网民的人口特征属性、浏览行为、历史行径等多种交叉维度的详细数据进行考量和计算，整理出大量的消费者数据，作为营销决策的重要依据来源，具有极高的参考价值和应用价值。

RTB 凭借大数据的模式，真正做到了用数据说话，与用户真切沟通。运用真实数据，分析真实数据，通过强大的整理、分析、计算功能使数据真正变为营销分析的有效依据。RTB 模型就是从实际出发，根据用户行为量身为其打造的广告环境，其科学的方法比传统互联网广告方式有着更多的长处。

目前全球网民数已经突破 30 亿。而根据中国互联网络信息中心（CNNIC）发布的第 37 次《中国互联网络发展状况统计报告》，截至 2015 年 12 月月底，中国网民数量达到 6.88 亿。数量庞大的网民反映了现代人生活的习惯，“上班在互联网上，下班在移动互联网上”的状态正是现代人生活的真实写照。消费者通过互联网建立起来一个全球的社区生活，任何人留下的痕迹都是体现个人商业价值的重要参考。

从互联网广告领域看，目前国内在线实时网络广告已经形成了完整的生态环境，广告位提供方、广告交易平台、广告投放商和广告主等整个广告价值链的参与方在不断丰富和完善。从图 5-5 中可以看到整个生态环境的构成关系。

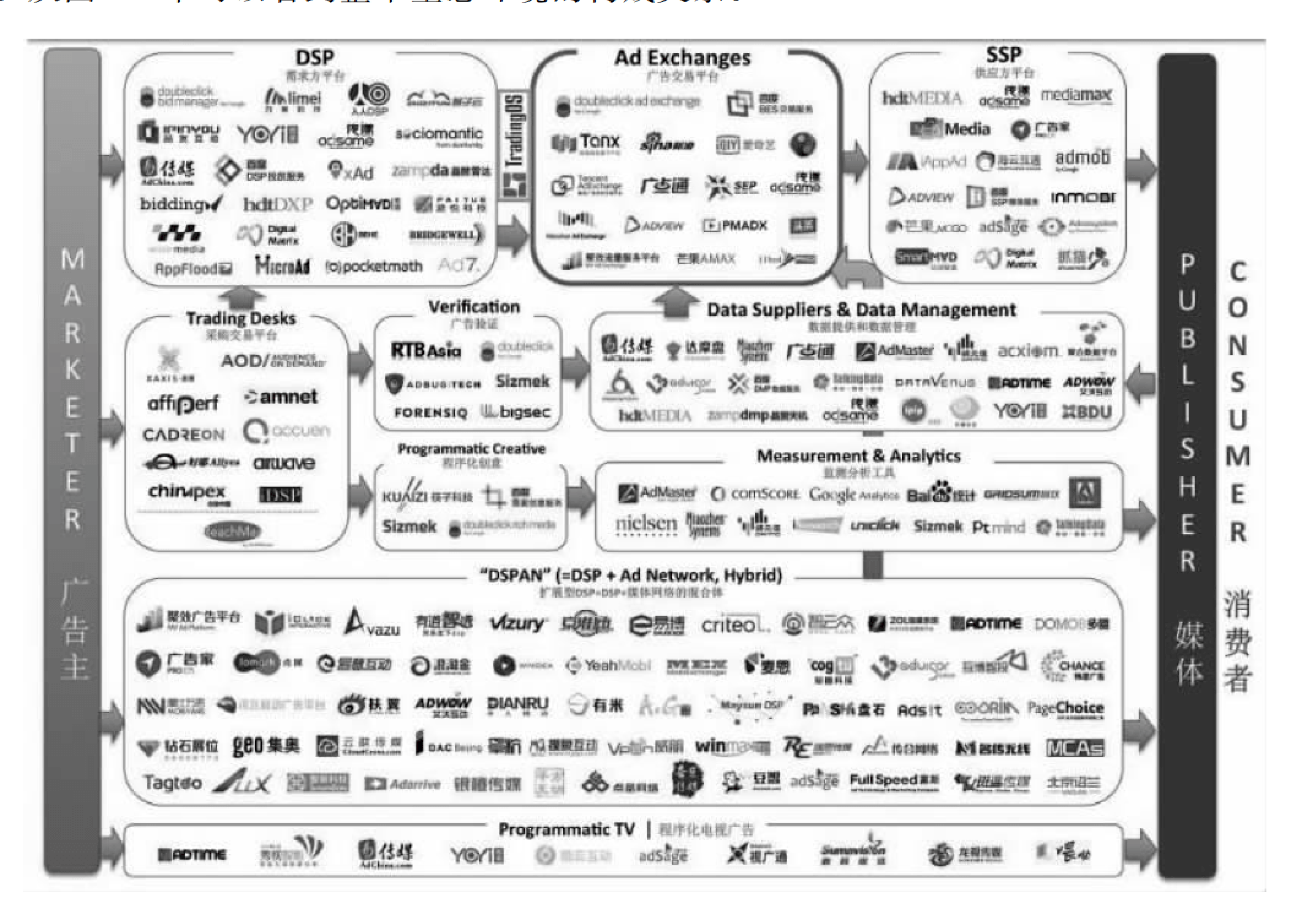


图 5-5　中国程序化广告技术生态图

（数据来源：RTBChina）

1. RTB 模式的所需技术平台构成

1）DSP

DSP（Demand-Side Platform，需求方平台）就是为有广告投放需求的广告主和广告代理机构而设立的平台。需求方平台汇集了各种广告交易平台、广告网络、供应方平台，甚至媒体的库存，允许广告客户和广告机构更方便地访问，以及更有效地购买广告库存。这个平台存在的意义在于帮助广告主在互联网上的众多媒体及媒体形式下进行广告投放，各家 DSP 都会通过各种技术手段做出一个系统，便于广告主在该平台上投放和管理活动。在竞价方面各 DSP 都会有独特的逻辑和算法，在人群的定义方面，DSP 会在系统中将每个 Cookie 上打上标签，通过标签组合的方式定义相关人群，帮助广告主找到相应的目标受众。

DSP 从投放效果来看，最大的特点在于可以帮助广告主实现智能化、合理化地购买广告流量。通过 DSP，广告主可以有效找到目标受众，通过数据分析进行合理的出价及投放，大大地提高了效率。在 RTB 模式中，DSP 需要有一个强大的基础设施和资源来为广告主实现在广告交易平台中的迅速竞价。因为在实现一次竞价的过程中，DSP 平台只有几十毫秒的时间，并且在这段时间内，DSP 还需要分析很多的数据，最终还要上传给网络广告交易平台，如果在这几十毫秒内 DSP 不能完成这些程序，网络广告交易平台将被认为是不能接受的投标响应超时，广告主的广告就无法被展现出来。在 DSP 系统中，会以用户基础为主，

来完成DSP竞价函数的确定。DSP采用基于数据的用户定向技术，将网络交易平台传输来的每一次曝光机会进行详细的数据分析，通过严密的逻辑关系和算法，最终决定是否对这次展现机会进行竞价展示。

这一平台拥有强大的数据资源，并且可以通过科学的计算方法帮助广告主选择所需要的数据，从而可以避免广告主的投资浪费，从理论上实现广告主预算零浪费的目的。因为这些数据的分析直接影响着广告主投放的广告效果，广告主选择数据做出定价时所需的就是DSP。在国内的主要DSP有品友、聚效、亿玛、新数网络、悠易互通、易传媒、广点通等。

2）SSP

供应方即指广告位的提供方媒体，同广告主在购买广告时需要DSP一样，互联网媒体在卖广告的时候也需要一个管理平台，也就是SSP（Sell-Side Platform，供应方平台）。SSP让媒体主也介入到广告交易中，实现相对精准的人群定向，智能地管理和帮助媒体整理和储存广告资源并进行合理的投放、优化。SSP的价值在于帮助网络媒体实现其广告资源优化，提高其广告资源整合价值，从而整体提升媒体成本效率。

供应方平台能够让数字广告发行商和互联网媒体的广告库存鲜活起来。通过SSP，数字广告发行商及互联网媒体都可将自己的剩余能力有效地利用起来。因为通过RTB模式投放广告，广告的位置已经没有找到这个人的相关属性重要了，在RTB模式中，广告位的优劣并不能绝对地说明投放的效果好坏。也就是说，在RTB模式中，互联网媒体将自己的广告位通过SSP进行售卖，从而达到更高的经济效益。现在，国内的主要SSP有品友和好耶等。

3）DMP

DMP（Data-Management Platform，数据管理平台）作为数据提供及管理的平台，在实时竞价广告中也至关重要。我们知道，在大数据背景下，RTB能够实现的依据就是基于海量数据的分析计算，DMP的作用就是管理一切与用户数据有关的部分，将所有与有关目标受众的数据打通，实现对目标受众的重新定向。

数据服务商可以基于网上数据基础，运用大数据工具，对互联网用户的基本数据及相关的行为进行分析，刻画用户特征，完成用户画像，进而做到每个用户特征的数据化，建立完善的用户标签系统。数据服务商可以获取得到全网的用户数据，具有覆盖面广、种类多、体量大、质量高等优点，结合实时的采集分析系统，进行海量的并发运算，实现多维度的数据处理，提供多维度的用户标签，包括用户画像标签、行为标签、兴趣标签、实时标签和历史标签等标签数据。

数据管理平台，通过广告交易平台（Ad-Exchange）、广告供应方平台（SSP）和需求方平台（DSP），帮助广告行供应商和广告主，完成个性化推荐、实时竞价（RTB）以及广告投放等多种服务，最终使得媒体推出的广告更精准，广告转化率更高。

市场上的DMP整体可以分为第一方DMP和第三方DMP。所谓第一方DMP，是广告主的私有DMP，收集整合的是广告主的第一方数据，包括广告数据、官网数据、EDM数据、CRM数据等，广告主拥有系统的唯一控制权和使用权。而所谓第三方DMP，控制权和使用权一般归DMP运营商所有，其中收集整合的数据不属于任何一个广告主独有，对于广告主来说属于第三方数据。目前国内独立DMP的发展还比较薄弱，因而有很多DSP会做自己私有的DMP以此来收集更多的Cookies积累数据，进而吸引更多的广告主。目前国内主要

的 DMP 有百度 DMP、阿里妈妈、广点通、易传媒、秒针系统、安客诚、缔元信、Admaster 等。

4) Ad Exchange

在 RTB 模式中,网络广告交易平台又起着特殊的作用。在网络广告交易平台中,首先是将媒体的广告位通过数据分析的方式进行整合分类。网络广告交易平台在 RTB 模式运行中发挥了更多的特效。在该模式下,网络广告交易平台整合在线媒体广告位资源,将其分析并分类,然后广告主可以通过自己的需求,在平台中选择自己需要的数据进行竞价,最终,网络广告交易平台将广告位售卖给出价最高的广告主。而这种数据的购买,其实是广告主在选择数据的同时,就是在选择一定量的具有这种数据的独立访客(Unique Visitor,UV),即独立 IP 地址。可以理解为广告主在 RTB 模式中购买的是一个具有特定数据的 UV。Ad Exchange 将会把某一特定用户已经单击某一特定网站的广告位公布出来,交由 DSP 为广告主做出竞价决定,AdExchange 在这个广告位投放广告的广告主将是出价最高的那个。RTB 模式的运行是在广告交易平台上实现的,所以 RTB 模式产业链中最主要的也就是广告交易平台。现在,主要的广告交易平台有谷歌 DoubleClick、阿里巴巴的 Tanx、腾讯 tae、Sohu、Sina 等。

2. RTB 模式的运作流程

RTB 模式的具体过程是,当用户单击一个网页,网页中实时竞价的“按钮”就被打开了,并在 100ms 内迅速完成。

数据服务商的 DMP 在用户之前浏览网页时早已记录下其浏览行为,DSP 在投放前期做了充足的准备,例如,在广告主网站或者活动页面加上代码,在投放的前两周时间里,收集来过广告主网站或者活动页面的 Cookies,并记录下来。DSP 将用户浏览网页的行为进行了精心的处理,以让广告主容易理解的方式呈现出来,广告主了解了该用户的特征,分析出了相关信息,如该用户的性别、年龄、爱好、之前都浏览过哪些类别的网站等信息。广告主选择是否竞价,结果就会在瞬间知晓,出价最高的广告主将获得这次展现广告的机会,而竞得的价格将是出价第二名的竞价价格再加一分钱。

当投放正式开始时,各环节在 RTB 模式运作过程中紧密合作。SSP 将广告位资源集中在一起,便于管理,每当有广告位和 Cookie 出现就会向 Ad Exchange 发出讯号,告知其广告位置有空缺,可以竞价。Ad Exchange 每当出现一个展示机会都会向 DSP 发出信息,DSP 除了应用以往的投放数据和对人群属性的分类数据,在系统中自动帮助广告主竞价。DMP 在此过程中可以帮助 DSP 丰富人群数据,因为 DMP 专门做数据的整理工作,会帮助 DSP 更加合理地做出出价判断。

用户浏览任何网页时,随着页面加载几乎都会出现广告。按照传统做法,如果同时在两家门户网站的运动频道页购买了广告位,就不得不为两个网页重合的用户买单。实时竞价希望通过让营销人员购买目标受众群而改变游戏规则。因此,竞价胜出的营销人,最后展示的广告通常是“定制”的——在正确的时间、地点展现给正确的用户,这就是为什么在晴天男性用户可能看到更多的敞篷车广告。因为更具相关性,实时竞价广告容易吸引用户的兴趣。如果广告主知道某用户曾经访问过他们的网页,就会在该用户访问时用数字标记该用户的计算机(Cookie)。广告主会在随后的广告位竞价中付出更高的价格,再次接触这样的用户。

图 5-6 所示为程序化广告信息交易核心流程图。

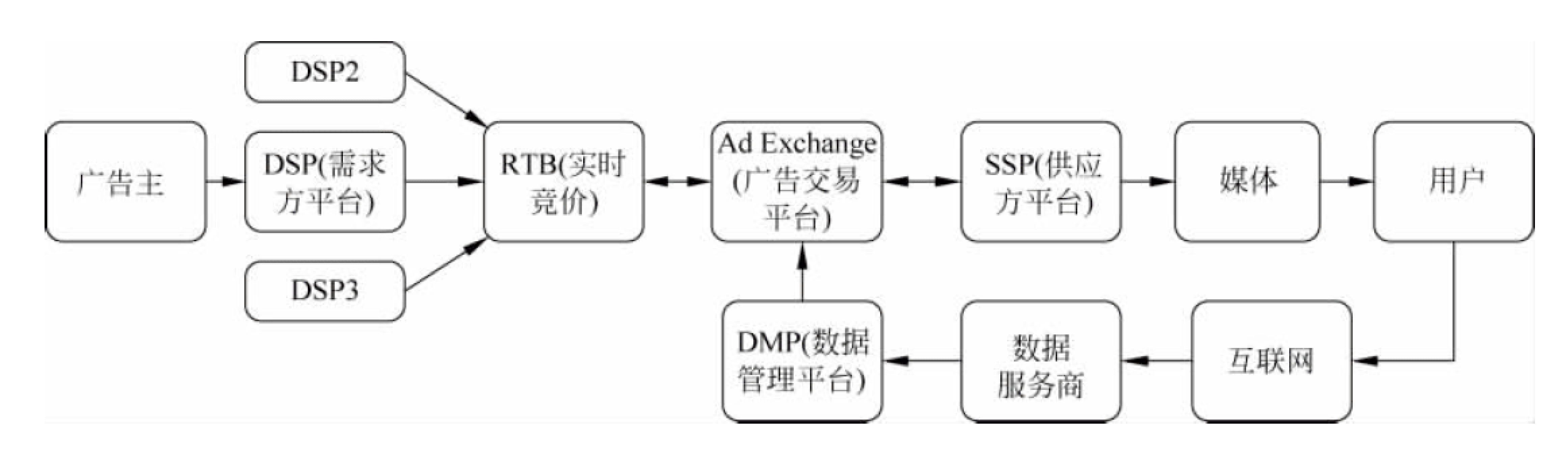


图 5-6　程序化广告信息交易核心流程图

5.2.3　RTB 投放工作内容

1. SSP 的媒体资源管理

当用户浏览页面时，SSP 的媒体资源管理获得用户浏览信息，将广告资源挂起，并对 Ad Exchange 广告交易平台发出用户访问信号，DMP 开始对来访用户进行行为分析。SSP 需要设定广告展现的个数与轮播次数，根据用户的注意力热图以及关键字出现的位置合理分配广告展示的位置，当然，也要综合考虑展示广告的大小。

2. DMP 用户行为分析

当 Ad Exchange 收到用户访问信号时，DMP 会针对用户的浏览信息进行分析，这种基于大数据背景下的数据挖掘需要容纳海量的信息，DMP 在法律允许的范围内，统一处理用户信息，使得第三方平台具有同一规格的分析指标，各广告主处理起来更具相对性和参考性。DMP 分析用户行为时可以从两个方向进行分析，一是从用户的实时需求入手，对用户搜索的关键字进行最优匹配分析；二是从用户的内在行为挖掘入手，对用户历史行为进行广告推荐。因此，DMP 综合考虑了用户过去的、现在的和未来潜在的广告项目，且分别从用户自身和其他用户的角度对每位用户进行细化剖析，有助于增加用户与 DSP 匹配的广度，在一定程度上增加曝光广告的多样性。

大数据的应用主要在这个环节发生，通过数据挖掘和数据分析确定相应用户，打通广告与用户精准匹配。

3. DSP 竞价策略分析

DSP 竞价策略分为两种，一种是固定竞价策略，一种是智能竞价策略。

固定竞价策略：即只要 Ad Exchange 发出竞价邀请，一律进行竞价。竞价金额可以通过以下几种方式实现：固定价格、周期价格等。设定这种竞价策略的目的主要是为了服务于中小企业或刚起步企业的短期竞拍投放，这在一定程度上缓解了部分数据冷启动的问题。

智能竞价策略：广告主在 DSP 端会参考 DMP 分析得到的用户指标来制定自己的竞价价格，而价格的制定不仅要考虑到用户的因素，也必须要明确广告项目的投放情况，进行风险预测和效果评估，以此来确定最合适的价格。对于智能竞价，各 DSP 广告主有自己的数据分析中心，通过 DMP 传递过来的某用户的 IP 可以识别该用户是本网站的新用户还是回访用户。若是新用户，则根据 DMP 传递过来的关键字匹配排名、本网站推荐广告排名、关联度排名、用户历史活跃度。

4. Ad Exchange 交易平台

Ad Exchange 竞价排名公式，可参考淘宝直通车，综合排名得分如下：

竞价排名得分＝质量得分×出价

在 RTB 中，质量得分为关键字匹配度、协同推荐项目评分、关联规则支持度阈值、用户历史活跃度这 4 项的加权求和。

实际扣费的价格如下：

广告项目实际扣费(元)＝下一名出价×下一名质量得分/本人质量得分＋a 元

实际扣费不会超过用户设定的出价。由于 RTB 增加了媒体维度，各媒体可分别制定自己的 a 值，一般 a 为 0.01。

5. 广告展示

Ad Exchange 交易平台负责完成交易报价，SSP 接到交易指令后，展示相关广告主的广告，并按照广告展现效果和约定的费用结算方式进行费用结算。同时交易平台也将收集用户与广告有关的数据，为后续广告需求积累数据。

5.2.4 RTB 应用场景示例

在现在的互联网实时在线广告市场上，通过对用户数据进行有效的信息分析，可以进一步实现高效、精准的广告投放，实现广告在适当的时间精确地一对一传递到目标用户面前。

简单来说，该广告的商业模式就是让广告主付出合理的价格，在“正确”的时候，使“正确”的广告显示在“正确”的目标受众面前。

某用户在某电商网站上近期浏览过一双运动鞋的产品，DMP 通过对该用户数据的反向追踪，结合数据库历史数据进行匹配和精准识别，发现该用户为体育达人，则将广告推荐内容调整为运动类广告、运动鞋等，将数据反馈到 SSP。SSP 将该用户信息发布在 RTB 上，在 DSP 上的一系列广告商在 RTB 上展开竞价，某广告商最后获得了广告展现权，结果是当该用户下次访问有相应 SSP 广告展示位的媒体时，就能收到该广告商精准的关于运动鞋的广告，如图 5-7 所示。

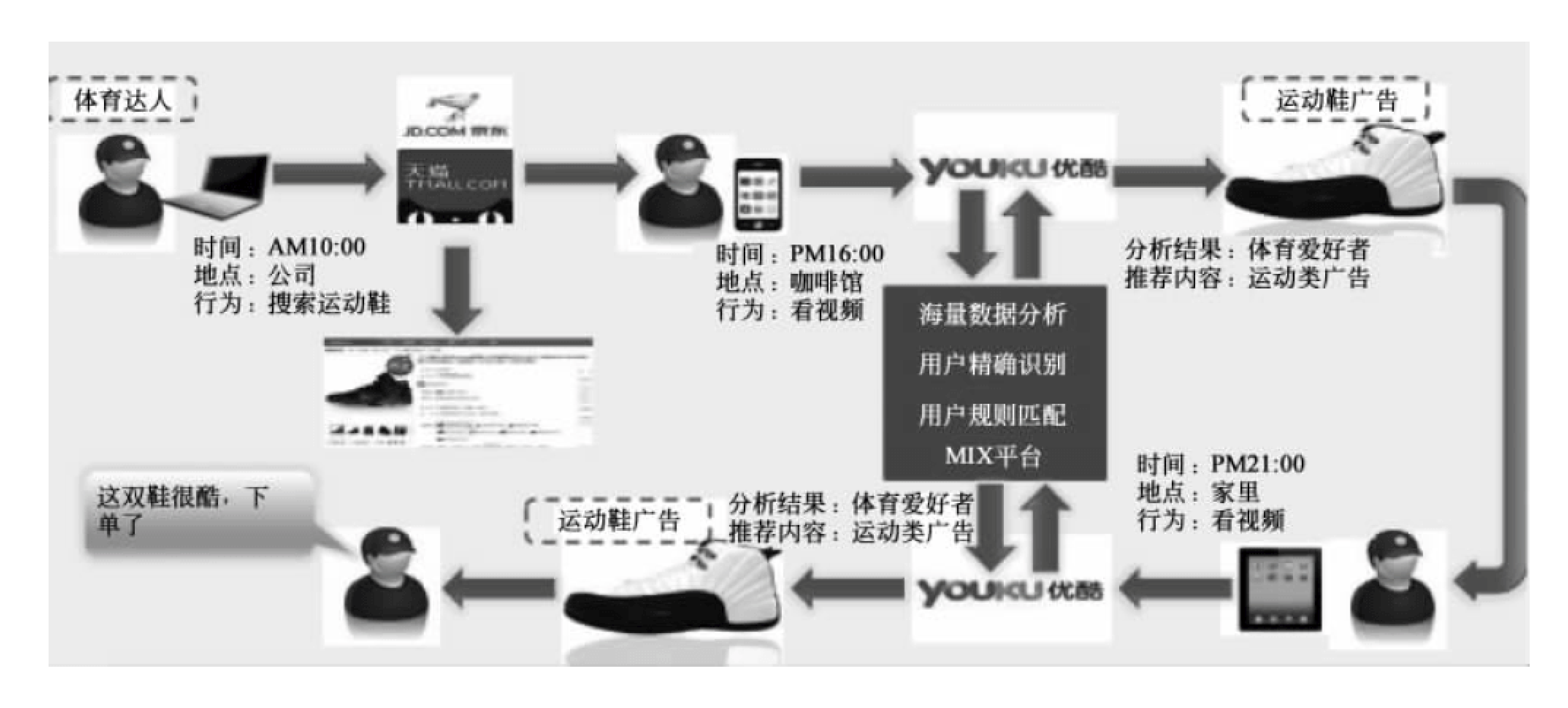


图 5-7 应用场景示意图

从可获得数据的内容来看，互联网服务提供商还可以通过对用户的基本特征、网站浏览行为、下载行为、增值业务行为等进行分析，了解用户的各项行为特征，从而为广告产业客户提供相应的服务。

以常见的手机用户的行为分析指标为例，主要包括如表5-1所示的几类。

表5-1　手机用户的行为分析指标

用户基本特征	性别、年龄、籍贯、手机类型
网站浏览行为	常用网站、浏览内容、手机购物
下载行为(内容)	音乐类别、软件类别、书籍类别、游戏类别
增值业务行为	短信、流量、咨询、话费、导航

互联网服务提供商可以通过对用户的相关网络行为、业务使用行为的数据分析，发现特定用户群体的业务使用情况或特定行为与特定业务之间的使用关联情况，进而可以对用户提供精准的增值业务服务推荐，如对客户群体进行细分，针对其短期行为提供相应的增值业务推荐。

其他场景下，如银行金融机构借助海量数据，创建数据分析挖掘为核心的精准营销平台，实现银行、信用卡客户、品牌互相联动的全方位、精准服务，为特定用户智能化地推荐有价值的业务和相关金融服务；电商平台通过用户的数据进行分析挖掘，确定每个用户的具体数据标签，根据用户的实时需求做到精准的个性化商品推荐。

大数据与云时代的到来，为海量数据存储、处理提供了强大的技术驱动与支撑。基于精准营销的RTB广告模式借助大数据的强大优势，将大数据技术、数据挖掘和预测应用到展示广告上，通过对海量的结构化和非结构化的数据进行分析处理，运用真实数据，分析真实数据，通过强大的整理、分析、计算功能使数据真正变为营销分析的有效依据。在原有的以覆盖面为核心的广告模式的基础上采取精准推荐、实时竞价的方式实现了广告效果的优化提升。

大数据技术的应用为广告效果的精准推荐、广告交易的实时处理提供了很好的解决方案。但目前受限于处理技术和分析数据的能力，广告推荐、交易并未实现精确的实时交易，从数据发现到处理匹配、推送请求再到交易平台完成处理，还是有一定的时间延迟，且在精准程度上也尚存较大的局限性，有待于对分析处理模型和数据计算能力进一步的探索。相信随着大数据计算的发展，在精准营销广告的推荐上将会有更大的提升，真正能够实现在适当的时间将适当的广告推荐给适当的用户的场景，而此时也能够更好地实现广告主、交易平台和用户之间三方共赢的局面。

RTB得以发展成熟缘于互联网广告与互联网数据的紧密结合，没有这个结合就没有RTB的今天。在大数据应用中，RTB能够异军突起是因为互联网广告与互联网数据同在数据领域之内，专业相通，它们的结合顺理成章、顺水推舟、顺势而成。但一般来讲，各行各业与数据应用的专业之间相距甚远，并不存在这样的近水楼台。如欲在各行各业发展大数据应用，首先需要各行各业的各级各类业务人员或领导能提出数据应用的需求，不能提出需求，便不会有后面数据应用的出现。

大数据应用需求隐藏在各行各业的各类业务之中，需要千千万万的业务人员把这些数量庞大且价值与作用巨大的需求发掘出来，别人是不可能替代他们的。发展大数据或数据分析的唯一途径就是通过全行业海量的恰当培训，逐渐提升各级各类业务人员的数据意识与素质，普及面越大越好，培训的水平越“高”越好。发展大数据或数据分析应用是一个过程，这个过程的长短取决于近乎全民的数据意识与素质提升的速度，这是一个巨大的工程，不可能一蹴而就。一万年太久，只争朝夕。



第 6 章　阿里的数据王国

2009 年 9 月 10 日，阿里巴巴十周年庆典上，阿里巴巴云计算团队以独立身份出现，命名为“阿里云”的子公司正式成立。在 2015 云栖大会上，阿里云发布全新品牌口号及品牌广告——“为了无法计算的价值”(Creating value beyond computing)，深入地阐释阿里云的品牌定位及品牌价值①。

2012 年 7 月，为挖掘大数据的价值，阿里巴巴集团在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台——“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。马云强调：“假如我们有一个数据预报台，就像为企业装上了一个 GPS 和雷达，你们出海将会更有把握。”因此，阿里巴巴集团希望通过分享和挖掘海量数据，为国家和中小企业提供价值。此举是国内企业最早把大数据提升到企业管理层高度的一次重大里程碑。阿里巴巴也是最早提出通过数据进行企业数据化运营的企业。如图 6-1 所示为阿里大数据体系。



图 6-1　阿里大数据体系

在传统认知中，“计算”一词对于大多数人而言太过遥远和冰冷，那是必须花费力气去破解的代码世界，与日常生活的交集看起来是那么微乎其微。然而，阿里云认为，计算的终极意义是发挥数字的力量，去解决问题、创造价值，让数字不止于数字，赋予数字以人的喜怒哀乐。6 年的光阴更见证了计算对生活、对社会、对每一个普通人产生的潜移默化的影响，那是科技理性与人文感性的精彩碰撞，在和谐之中共享无法被衡量的价值。

① 阿里云推全新 Slogan：为了无法计算的价值[OL]. 新浪科技，2015. 10. 14，http://tech. sina. com. cn/it/2015-10-14/doc-ifxirmqc5116488. shtml.

阿里云的服务群体中，活跃着微博、知乎、魅族、锤子科技、小咖秀等一大批明星互联网公司。在天猫双 11 全球狂欢节、12306 春运购票等极富挑战的应用场景中，阿里云保持着良好的运行纪录。此外，阿里云广泛在金融、交通、基因、医疗、气象等领域输出一站式的大数据解决方案。阿里构建起了一个通过多维数据来描绘用户画像的数据王国，每个用户既是这个数据王国中的数据生产者，也是数据应用的服务对象。

6.1 "滴滴打车"助市民出行无忧

城市打车场景是"衣食住行用"5 大刚需领域之一，但在现实社会中一直存在乘客打车难、出租车空驶时间长等难题，而"快的打车"App 利用移动互联网专注于产品与大数据，为乘客和司机提供更好的出行解决方案，解决中国大众出行需求。"快的打车"的出现已经改变了传统模式下用户出行的行为习惯，以及传统出租车行业的运营方式。目前，"快的打车"覆盖国内 360 个城市(含中国香港)，据艾瑞最新统计，"快的打车"用户覆盖比例超过 60%，用户数量过亿，是国内最大的出租车叫车软件。

2012 年 5 月"快的打车"成立，2013 年 11 月，"快的打车"并购"大黄蜂打车"，2014 年 7 月正式推出服务于中高端用户的用车品牌"一号专车"，2015 年 2 月"快的打车"与"滴滴打车"合并。现在的"快的打车"已经从最初的打车软件成长为全国最大的移动出行平台，并与支付宝、高德地图、铁路管家、国航、如家连锁酒店等各类出行场景深度合作，并以数据驱动为重点，进一步提升出租车、专车运营效率，丰富用户出行体验，在这个万亿级的市场快速成长壮大。

根据艾瑞咨询《2016 年中国移动端出行服务市场研究报告》显示，截止到 2015 年年底，中国移动端出行服务用户乘客数量总计接近四亿，滴滴专车(快车)用户覆盖数量占比高达 88.4%，同时在中国专车(快车)移动端出行服务行业中，滴滴专车(快车)日均订单量占比达到 84.1%，滴滴出行已成为中国服务内容最丰富的移动出行 App。[①]

图 6-2 预测了从 2012—2027 年整个出行方式占比，第一个是步行，第二个是自驾，第三个是一些公共交通，大家坐的地铁、公交等。这张图最典型的变化就是自驾出行与公共交通出行，在 2027 年时，达到 97%左右。这就带来两个问题，第一，自驾场景怎么解决，第二，我如果坐公共交通，比如出租车、大巴、火车，如何能最佳匹配我的需求。"分秒"之间的多轮筛选，数据完成的用户画像系统，人们点滴的打车轨迹正在汇聚成一个全新的商业生态。

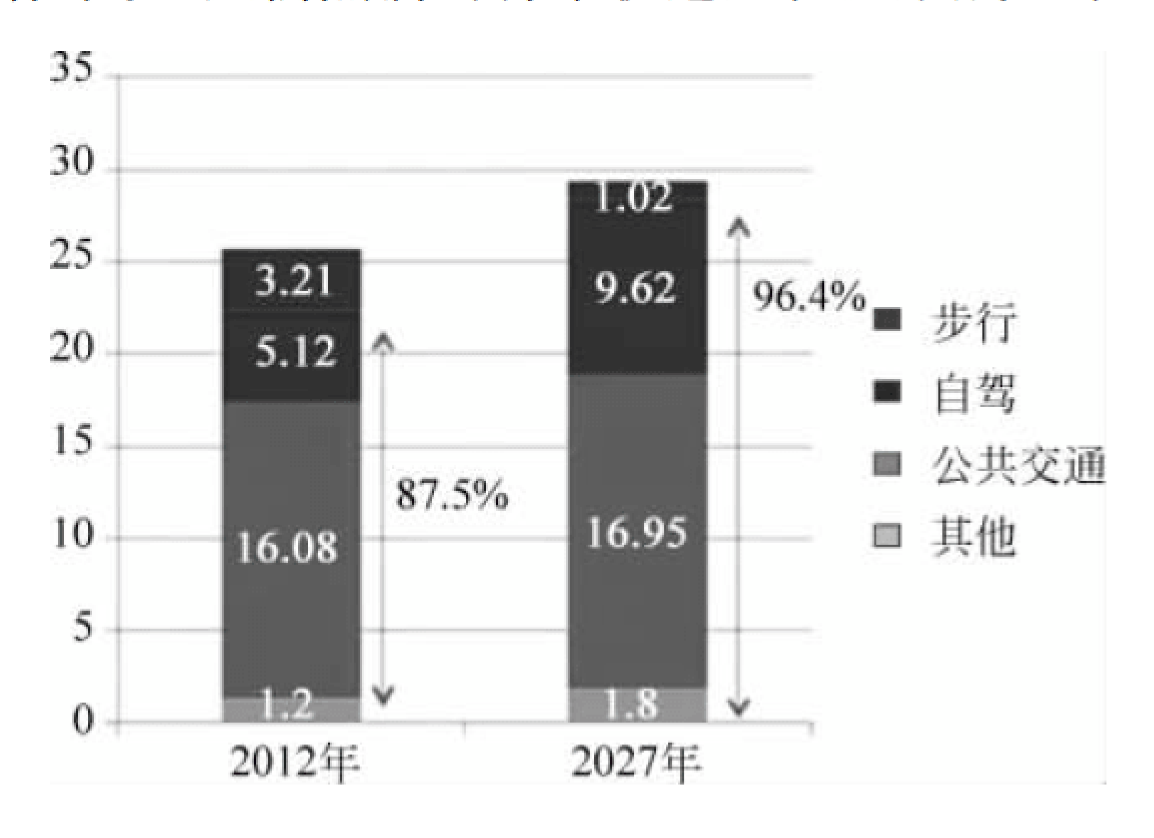


图 6-2　出行方式趋势图

① 2016 年中国移动端出行服务市场研究报告[R]. 艾瑞咨询，2016.

如图 6-3 所示为 O2O 供需智能匹配。



图 6-3　O2O 供需智能匹配

6.1.1　典型案例

“互联网＋交通”与传统交通的最大不同在于创建了基于互联网全网信息的多边供需平台，并利用移动终端不间断地收集供需数据（人、车、货的地理空间等交通配套信息）、调度供需双方并实时交流，采用大数据进行实时分析匹配，采用云计算以质优价廉的计算能力全天候支撑“互联网＋交通”供需平台上全民出行、全国货运的智能服务。打车软件承载的是人、时间、空间多维度结合的生活场景，个性化推荐更加投其所好，贴近实际，转化率会高很多，如图 6-4 所示。

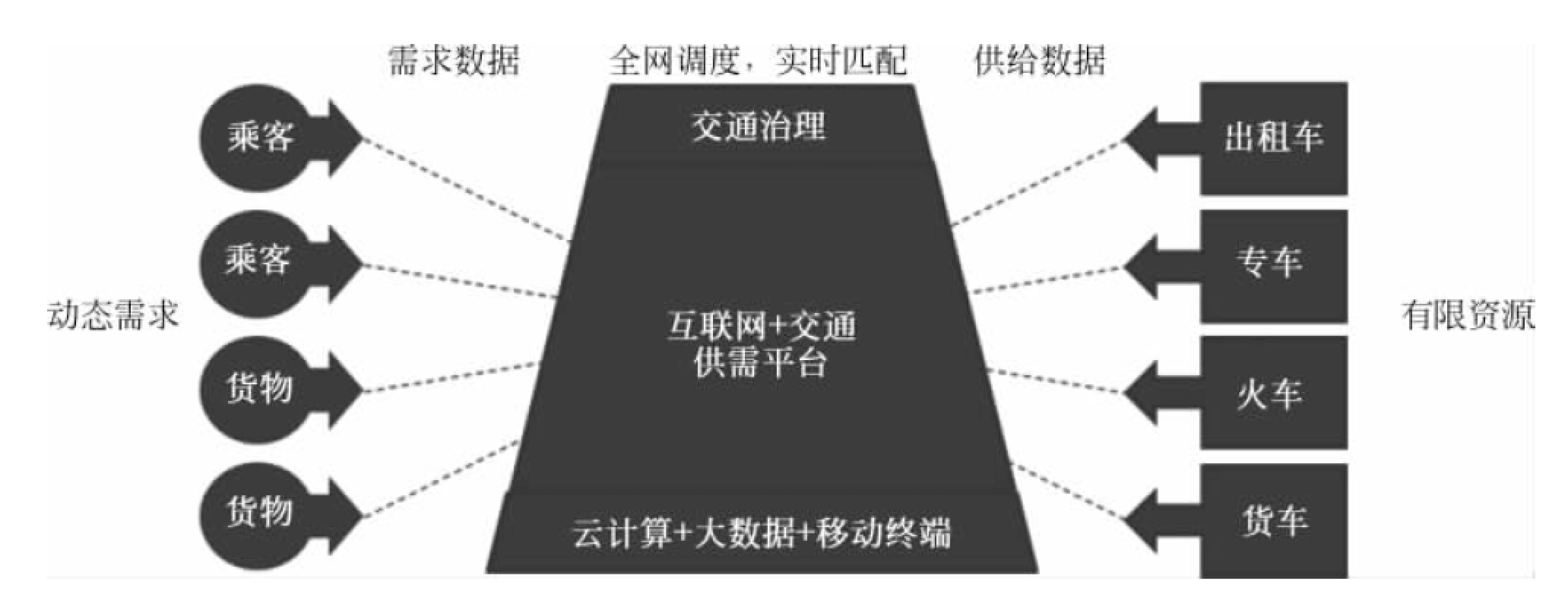


图 6-4　互联网＋交通供需模型

【例】　周五晚上 6 点 40 分，李菲（化名）在离家不到 3km 的地方，用打车软件叫了一辆出租车，在不到 1min 的时间、系统通知了附近 43 辆出租车之后显示被抢单，与此同时，李菲的手机上收到一条短信：“我们额外支付了司机 11 元，这部分费用由土豪快的买单。”

这次打车给李菲带来的愉悦感可想而知：之前望眼欲穿的苦等，现在则分秒可得。不过李菲或许不知道的是，从她按下快的界面的叫车键到系统启动用车通知“分秒”之间，快的后台已经完成了多轮筛选：根据用户画像和用车需求，匹配位置合适的出租车，再结合实时的地理位置和运能状况确定给后者的补贴金额，这些计算都是在毫秒内实现。甚至在更早之前，快的已经根据她的历史打车的行为特点，将其划归到了“屌丝”的标签之下，由此她才频繁收到金额不小的代金券。

6.1.2　案例分析

1. 用户画像：屌丝和土豪的不同行为轨迹

快的"土豪式"补贴背后，其实也有着它自己的精打细算。就如快的公司技术副总裁朱磊对记者所说的，行业已经从粗暴的跑马圈地走入了精耕细作的时代，要花更少的钱获取更多的用户。

精准营销的前提是对用户的清晰认知。以简单的代金券发放为例，快的的历史数据呈现出两大类4种不同的消费习惯——代金券敏感型：发代金券才用，发代金券用得更多；代金券不敏感型：发与不发都用，发代金券也不用。在快的的用户画像系统中，上述4种群体会被分别冠以屌丝、普通、中产、土豪的标签。针对4类客群的运营策略也会全然不同，最直接的就是代金券的刺激频率以及刺激金额，而对"代金券"免疫的土豪群体，则更多地需要在服务上做文章。

而在实际场景中，影响乘客对应用软件的使用黏度的因素要远比代金券复杂得多，在这种情况下，快的对用户的"贴身跟踪"就能及时发现薄弱环节，因此从用户打开软件到退出使用，其间的每一步情况都被快的记录在案，如哪一天退出的，哪一步退出的，退出之后"跳转"到什么软件等。

据此，快的也实现了用户另外一个纬度的归类，分清哪部分是忠实用户，哪部分可能是潜在的忠实用户，哪些则是已经流失的；更进一步来看流失的原因：因为代金券没有了流失？软件体验不好流失？还是等车时间太长而流失？这些都是下一步精准营销的依据。

而对于快的而言，用户分析不仅是针对乘客，也包括司机、出租车公司的所有相关方。尽管基础信息大同小异，都包括人的基本信息、信用、行为信息等；也有一些通用的刺激手法，比如积分、礼物等。不过，不同的用户画像就对应了不同的刺激程度，而结合不同的场景，还有许多特殊的营销安排。

杭州市场就是一个很典型的例子。基于司机的地理位置信息，快的发现每天中午或者是每天晚上10点以后，司机都会聚集在一些固定的地点，可能休息或者就餐。所以快的就会在这些场所提供一些工作餐或者是优惠食品，通过线下的活动来提升司机和快的的合作关系。

2. 产品生成的逻辑：更精确地匹配供需

维护好用户只是一个基础，最终目的是为了打通供需，生成更加优化的服务和产品。这也正是数据之于打车软件此类的O2O行业的重要性所在。"数据能解决一个核心问题，就是做供需双方的智能匹配。"朱磊说。

其实也很容易理解，公交、出租车、地铁都是对出行人群不同需求的对号入座，不过这样被朱磊称之为"粗暴式"的分类法应用起来效率低下，以一个司空见惯的打车场景为例，在路边拦车，可能许久都没有空车经过，或者是好不容易等到的车，司机问了地址之后还可能拒载——呈现一种杂乱无章的状态。

而在海量的数据基础之下，出行的需求被不断细分，而且是实时匹配。例如一个乘客下单之后，需求方的用户图像和需求同时被识别，结合供方的车辆条件和位置地图进行第一轮筛选，不过这个看似"正常"的订单却不一定符合实际，因为有一些订单发出来是司机不愿意

接的，比如高峰时段的拥挤路段，那么在这个时候就要进行订单评估和内部调节，结合历史数据制定一些刺激措施、叠加“乘客自行出的小费”来诱导司机，这样一个符合供需双方胃口的“合理”订单就生成了，下一步要做的就是实时调度，要考虑当时的交通情况、车的朝向、车速、附近是否有突发性事件等因素，选择最为优化的方案。

比如我要从A点到B点，发送单子后，向司机进行请求推送。这里面涉及一个简单问题，怎么告诉周围的司机最方便？最简单的策略是什么？有人说是距离，非常正确！这个逻辑非常简单而且特别高效，但是有没有问题？当然有问题！举一个负面例子，南京这个城市，中间有一条长江，有江南和江北，画一个圈计算离用户最近的司机，系统自认为聪明地把这个单子推给了江对面的司机，实际上这里根本没桥，直线距离5min，而实际过去的话要绕两个小时。这是典型的逻辑障碍，这该如何解决？这里就需要有完善的路径规划能力，对路况信息、路网信息、车辆信息，甚至是更细节的风向风速都要考虑。这就诞生了产品层面对大数据分析的最接地气的需求，行业内部叫智能订单推送系统。订单不再是这么简单的或者以比较笨的方法按照距离往外推送，而是结合了路况、路径规划、天气、车辆状况、车速、方向等一系列因素考虑。这套系统非常方便地实现了乘客需求与司机接单的智能匹配方案。

完成了以上的步骤之后，快的才会把用车需求和奖励方案推送给经过层层筛选之后的出租车，这样人们打车的成功率大大提升了，而且所用的时间更短。“这是以前所有的产品做不到的，因为不能洞悉消费者的心理。在大数据应用下，消费者和供给方能够省略中间环节直接议价，这是一个模式上的变革性的突破。”而最终海量的议价数据将提炼成为一种“商业情报”，来推动新的产品和新服务的推出，比如智能定价系统，以从机场到望京这一段司机不愿意接的单为例，可能70%的乘客额外加了20块钱，少数人加了30块钱，而有的只愿意加10块钱，那么系统整合分析以后会得出21元钱是一个更合适的议价，那么最终的定价可能消费者和司机双方都可以接受。

因此，以这样的逻辑推导生成的产品才更能有的放矢，因为其生成不是来自于企业对市场的臆断，而是直接提炼于供需双方的心理预期和真实需求。

“回程单”的产品创设就是一个很典型的例子。最初是快的的数据分析发现一个异常的数据现象，就是司机的抢单意愿率在某一个时点会骤然下滑，过一段时间又会反弹，日日如此。通过对这个特殊节点分析，快的得出一个司机运营的特殊场景，就是司机收工的时间，接下来就是针对性地解决，因为不管司机是交班还是回家，肯定有一个固定的方向——这一点可以通过历史数据分析出来。那么快的要做的就是把同样去往这个方向的乘客分配给对应的司机；这样做是否就一定见效？所以下一步就要评估效果，看回程单是否真正提高了司机的抢单意愿，确定之后才能作为常规产品推出。

“产品的细分应用场景将会越来越依赖于大数据分析，从数据中洞察需求与商机，再结合大数据提供应用解决方案，将变成未来产品迭代的常规运作模式之一。”朱磊说，这也是快的产品的生成逻辑。

3. 跨界的数据“火花”

尽管快的野心不小，想要构建一个全新的广告“生态”，不过这显然不是快的凭借一己之力所能实现的，必须借助于外部数据的导入，这恐怕也是大数据应用最基本的要求，那就是开放和共享。

与阿里和美团等的合作就实现了双方数据的相互补充，“他们缺乏出行数据，我们目前缺失的是用户的消费数据和信用数据。”朱磊说。在此基础上就可以共建用户画像体系：工作地点、家庭地点、消费情况、价格敏感度等。

在一个完整的用户图像下，广告推送就会更加精准。比如定位到一个北京用户打车去西单，在分析出其消费偏好的基础上，就可以针对性地发送特定商场特定店铺的某一类产品的优惠信息。“量身定做的实时实地的广告价值将远远超过传统广告盲目推送的方式。”一些针对节日的广告类型也会应运而生。以七夕节为例，就可以首先圈定跟节日消费相关的群体，提前两天推送花店信息，可以在节日当天直接送花上门，甚至可以制造一些小“浪漫”：或许可以设想一下当你的女朋友看到一辆豪车来接她下班时的惊喜，而车上还放着她喜欢的音乐，外加一束娇艳的玫瑰花。

完美的畅想还不得不面对现实中固有的一些问题，就像朱磊说的，一个来自于不同的行业标准和数据标准所带来的数据通用的难题，而即便在技术共享上不存在障碍，而协商机制的建立也将是一个漫长的“对话”。

“数据的价值评判每一家都是不一样的，那么就需要跨界的共赢机制的建立，这个在历史经验上是不存在的，只能去摸索磨合，这个过程肯定是痛苦的。”朱磊说。

4. 数据驱动模式的基础：技术投入

尽管还存在不少待解难题，如今开始把关注焦点转向数据驱动模式的快的，都已经与“补贴大战”时不可同日而语。因为任何的新兴业务，不论发展初期如何势如破竹，也必然要经过一个商业模式的探索和沉淀，否则最终会被“价格战”拖得精疲力竭。

“经过初期的野蛮生长之后，还想获得跨越式发展，就肯定需要在技术上的重点投入。”曾担任百度云计算主要负责人的朱磊，在快的带领的团队主要负责大数据体系、商业体系、基础架构与新业务等方向。

这个三四十人的团队在最初的三个月，经过了朱磊所说的“苦活、脏活、累活的痛苦历程”，进行了数据导入、清洗、存储、结构化等一系列最基础的处理，最终建成了快的的大数据体系。据朱磊介绍，目前扩建后团队的核心力量正在进行大数据 2.0 系统的研发。这套内部代号为“地平线系统”的大数据架构，克服了 1.0 系统中突出的数据数量与数据质量、处理速度之间的矛盾，实现了数据纯度、处理速度的跨越式升级。

这个“超级大脑”支撑了快的大数据应用所需要的所有基础数据，在此之上是支持产品、商业、运营商业化的团队，每个团队配备了 20 个人左右。这样的架构实际上避免了基础数据和应用数据之间的“污染”问题，比如一个需求场景形成了 A 的画像集合，其中结合 B 行业又会出现一个 AB 子集，应用到特殊的场景 C 之后又会形成一个同时满足 ABC 的集合。如果每次都从基础数据抽取，就很容易影响基础数据的稳定性。

清晰的数据架构对于“每秒(毫秒)都产生海量数据”的快的来说，重要性不言而喻。而今，数百台机器支撑着的快的大数据系统，在朱磊看来，它们就像是公司的“心脏”：业务规模越大，越是重要。

这种投入不是任何一个公司都能够负担的，却是每一个公司都应该及早想清楚的，“过早投入，对精力和资本消耗太大，不过如果之前缺乏考虑，后面就要做很多工作才能把之前错失的那些数据漏洞补回来。”对于早期一直争抢用户市场而忽略了数据应用的快的来说，这恐怕也是宝贵的“经验之谈”了。

总体来看，大数据的应用对快的业务发展产生了巨大价值。仅大数据产生智能订单这一个环节的优化，对打车成功率的提升幅度，相当于5亿巨额补贴同样的业务贡献效果。大数据前景的展望和一些挑战是很关键的，第一点就是大数据的基础架构的支持，未来的数据增长是十倍或者百倍，大数据架构能否熟练地运用起来，这是实施关键。第二点是生态圈的覆盖，谈到大数据就是不要把自己封闭，快的打车跟阿里合作或者跟高德或者百度地图合作，尽量开放，实现共赢，一方的核心价值是什么，另一方的核心价值是什么，大家做一个交换。第三点，跨行业合作如何基于数据统一去做。对于大数据面临的挑战，首先要考虑用户隐私，App能精确定位客户住哪、消费、信用等一些信息，这些信息泄露出来对人的影响非常大，这在最早设计的时候就要认真考虑并保护，快的要尽量模糊个体，不要定位到单个人，但可以定位到一群人，例如，不定位张三这个人是搞金融的，但是国贸这一群人是搞金融的人，这样可以有效地保护用户隐私。第四是整个数据架构的设计，因为之前讲的内容已经比较多，就不重复了。

6.1.3 “快的”的数据价值

快的打车的案例带给人们如下价值启示。

1. 商业价值

“快的打车”是一家1000人的公司，租用了400台云服务器，两年时间创造了30亿美元(公司估值)的奇迹，其关键在于充分发挥了移动互联网、大数据、云计算三者的赋能价值。智能手机的普及成为人类的身体器官延伸，便捷低成本地获知用户地点和出行模式，人口密度大的城市中被压抑的大众即时性出行需求在技术支撑下获得完全释放，应运而生的技术创新引爆了比美国更大的国内出行市场。

2. 数据价值

大数据技术的创新应用，对业务能够产生直接价值与推动力。大数据的开发与运用，必须深入到业务生产过程中，才能知道哪些内外部数据能够更好地服务最终用户，大数据应用先咨询设计、后开发实现。另外，企业的核心竞争力一定不是技术本身，而是基于“移动互联网＋大数据＋云计算”的生态价值，不同定位的企业共享数据、创新应用都能分得市场蛋糕。

3. 经济价值

快的打车有效地将出租车空驶率从40％以上降低到25％，该模式的经济价值体现在通过降低空驶率创造市场价值，空驶率每降低5％，每年出租车交易市场的规模将增加200亿元。采用打车软件，司机不用空驶获得足够订单量，挖掘剩余25％的出租车运能，创造出几百亿的新市场价值，而且无订单时节省大量汽油，平均一辆出租车一天的空驶里程大约90km，每天能省下几十块油费，又创造出几十亿元的新市场价值。另一方面，打车软件出现前，杭州出租车司机每年至少罢工一次，但有了快的打车，司机收入水平提高了，劳动强度也趋于合理，没人罢工了。

4. 社会价值

有数据显示，仅北京一个城市内，使用快的打车的出租车每年因此减少的二氧化碳排放量超过8万吨。快的打车在追求高效率的互联网企业中，关爱弱势群体，专门成立了针对弱势群体的“乐行联盟”，并开展了老人、孕妇免费接送等一系列活动，还推出了针对盲人及视

障群体的优化版软件，中央电视台曾对此进行了专题报道。

6.2　“聚划算”的智慧营销

阿里巴巴集团 COO 张勇说：“天猫‘品质’、淘宝‘万能’、聚划算‘活力’。”聚划算就是阿里巴巴和所有商家的“倚天剑”，曾任大淘宝 CEO 的张建锋说：“聚划算本身没有商户，它的商户是从两个平台里面选的，可能是淘宝可能是天猫。”聚划算最核心的价值就是它有非常强的规模化的能力，走量。销售模式是一个金字塔，顶层是聚划算这种团购模式。第二层次是一段时间之内销售很多商品，卖光为止。第三层次是平台模式，如天猫、京东。团购是在金字塔的塔尖，有非常强的出货能力。有些商品需要短时间内的出货能力，特别是生鲜类。可以说聚划算是整个集团业务里最顶尖的销售方法。但是天猫、淘宝、聚划算，是一个有机的配合的过程。聚划算网站是一个体验式营销平台，以聚新品、量贩团、商品团为主，致力于给买家提供极致性价比的商品，以有限的商品打造热门商品，支撑平均每日高 UV(Unique Visitor，独立访客)访问下的高成交额和高售罄率。

大数据有两个强项：一是挖掘好的，二是发现坏的。随着聚划算的高速发展，这两个问题衍生出越来越多的大数据解决方案。举例来说，好的方面，通过数据算法自动挖掘出好的招商商家与商品，根据兴趣点推荐给最合适的消费者，提高平台销量；坏的方面，对虚假订单进行监控和预报，识别秒杀器等作弊现象，更快更恰当地辅助运营处理纠纷。

聚划算平台整个业务流程可分为两部分，“招商”和“交易导购”。招商系统主要根据一定的策略和工作流程，选择出合适的商家、商品。交易导购负责在某一时间段内，进行前台展示，引导买家购买。聚划算一直在做数据化运营，利用大数据来决定招商与导购。聚划算的量特别大，备货、生产，都要一定的周期，这个周期导致很多时效性强的商品没有确定性就不敢去备货。大数据给这个平台带来了很多变化，能够在一定程度上降低卖家的风险，根据卖家申报的商品合适与否进行销售预测。大数据让聚划算知道，什么价位能卖出多少此类商品，这是传统企业没法知道的。所以聚划算员工会跟卖家做进一步的沟通：产品要做哪些改变，定在什么价位。机器给出选择，人工进行最终审核与确认。

6.2.1　商家端：数据化招商

在商家端，聚划算的整体运营发展路线，是从“人工”到“积累数据”，从“数据”到“模型”，从“模型”到“自动运营”的过程，经验积累数据，数据养起模型，运营越来越精准高效，历史数据判断哪个品类销售最好，预测未来销量，给出最优商品选择，审批人确认决策，商品团基于数据自动化运营，品牌团由数据支撑发挥运营创新性。

1. 聚划算的招商流程

如图 6-5 所示，聚划算的招商流程包括活动管理、品类规划、招商报名、报名审核、排期发布 5 个环节，却存在因为人员经验等差异化因素导致的审核质量问题、招商与导购效果问题、运营能力沉淀问题、小众品类发现速度问题、人工审核量太大等问题。所以在原有人工流程中，以“去运营化”作为突破口，加入了模型支撑，依靠接入各种数据模型，自动化执行品类规划、报名审核两个步骤，由活动管理提供模型输入(坑位数、品类范围等)，模型根据要求跑出数据，自动生成品类规划、发起商家/商品的邀约报名。

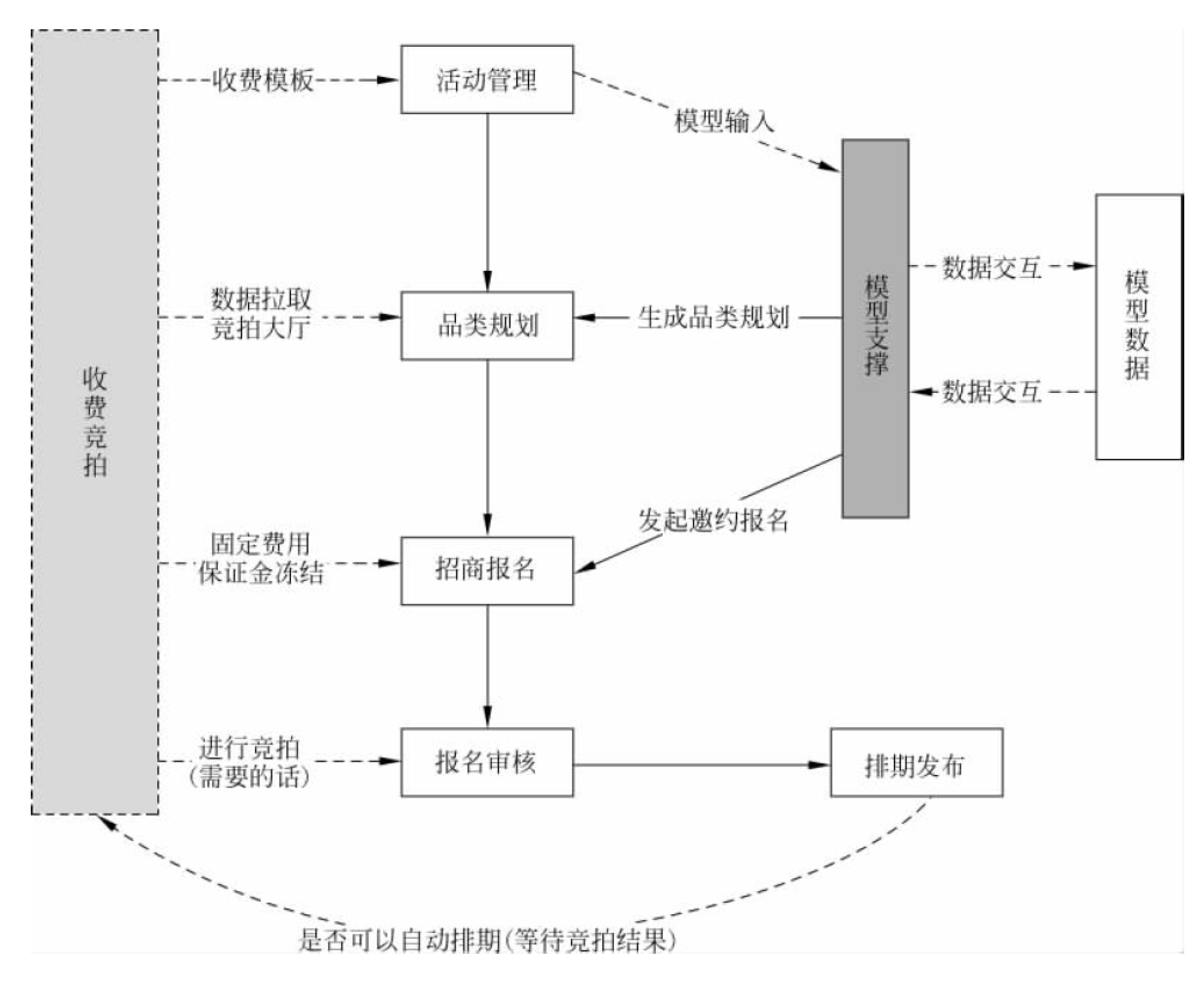


图 6-5 数据运营招商流程图

招商架构图如图 6-6 所示。

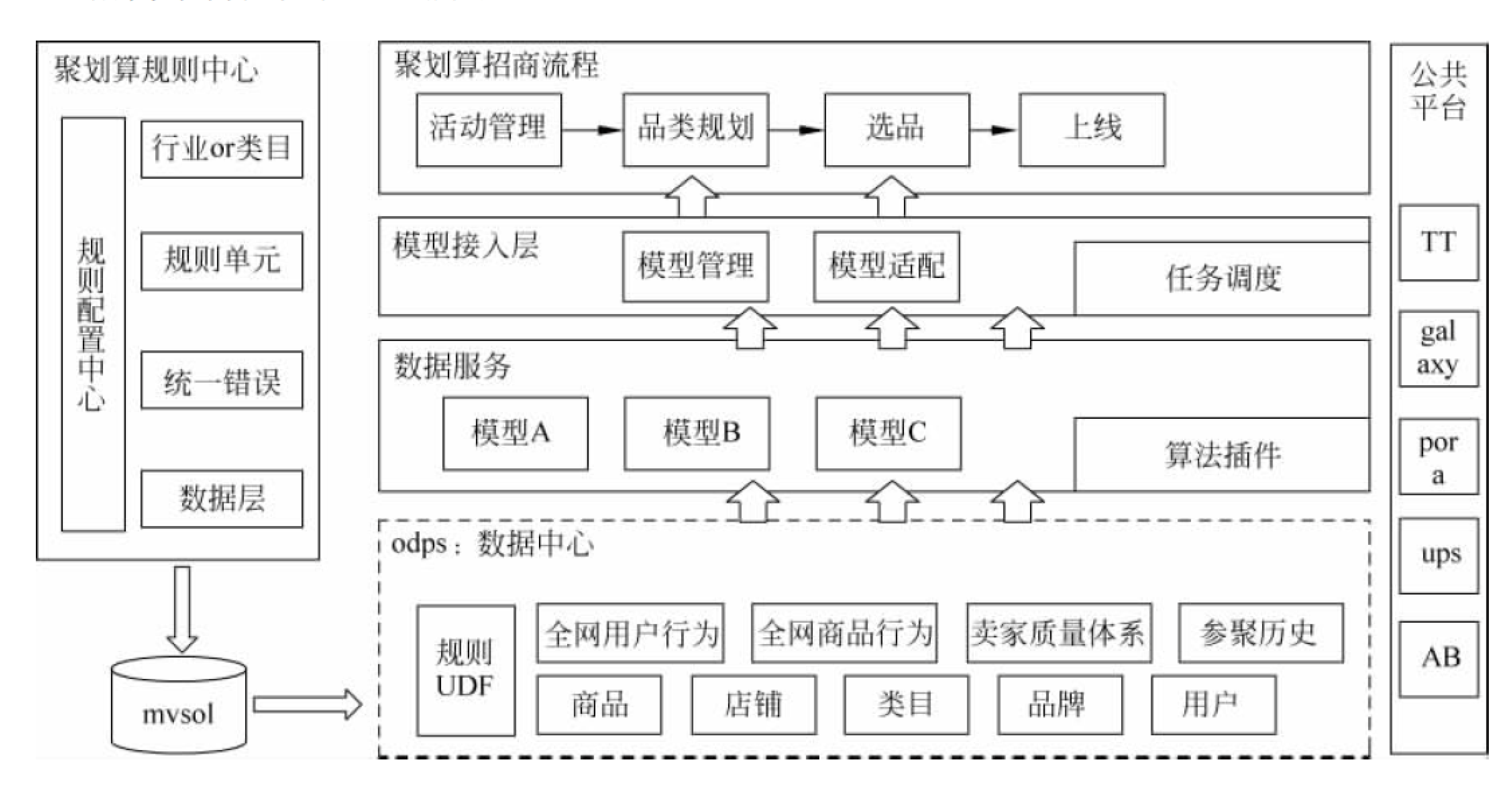


图 6-6 招商架构图

在整个招商流程平台上，规则引擎、数据模型、数据平台三个组成部分十分关键。

（1）规则引擎：是针对各行业类目的数据化运营“执法者”，贯穿整个参聚商品的生命

周期，既要融合准实时系统与离线选品模型，又要兼顾运营规则管理的灵活性、实时性、简单性，而且降低离线模型对规则的在线依赖。

（2）数据模型：是整个环节的核心、数据支撑的重点，关键数据环节包括数据类、工具类、算法类，数据类平台需要连接多个来源的数据，并提供基本数据查询、统计、展示和数据再加工等功能；工具类包括人群分析、市场分析、报表系统、人工选品等，需求是能够定制各种维度或组合的数据或数据报表；算法类包括主题挖掘、算法选品、潜客挖掘、主题创意、投放优化等，这类需求一般要根据业务的特点建模和优化，尽管需求不同，但底层的算法模型和基础数据具有共性。

（3）数据平台：提供完善的基础选品维度，以及快速整合数据资源，响应数据需求的能力是做优整个选品流程的基础，主要提供 4 种视角的基础维度特征。

① 商品维度特征：基础类特征（行业、类目等），浏览成交类特征（浏览、收藏、加购、购买及各项转化指标），运营服务类特征（上架时长、包邮退货服务、品牌授权等）。

② 卖家维度特征：基础类特征（主营类目、店铺类型、星级等），浏览成交类特征（浏览、收藏、加购、购买及各项转化指标，笔单价、客单价等），运营服务类特征（开店时长、熟客率、动销率、DSR 评分、有无消保等）。

③ 图片维度特征：基础类特征（宽高比、边框宽度），颜色显著类特征。

④ 行业维度特征：店铺在行业下的老客复购率等。

2. 数据应用

整个招商流程中有 4 个比较重要的数据接入点，包括品类规划、招商报名、报名审核、商品展示，其中，数据在品类规划和报名审核发挥较大作用，如图 6-7 所示。

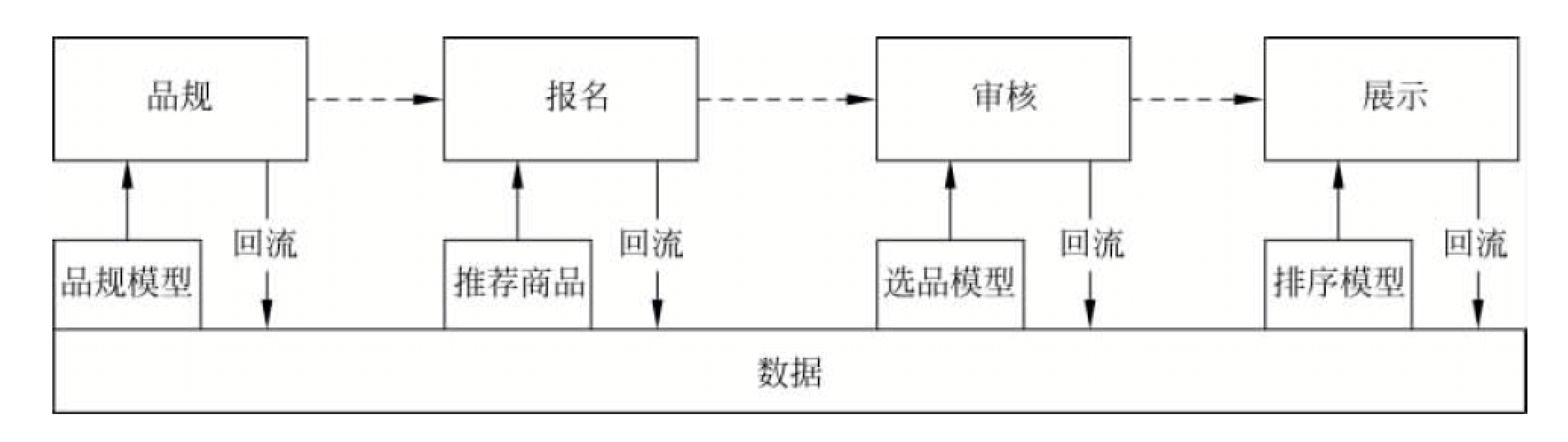


图 6-7 招商流程中的重要的数据接入点

（1）品类规划：是招商的根本，规划是否合理、品类范围大小直接影响招商质量、后续销售产出，通过大数据的积累，使用算法模型结合特定类目招商要求，保证品类、品牌丰富度等规则，生成品类规划，如图 6-8 所示。

（2）报名审核：分为商家审核和商品审核，是聚划算对商品质量控制的关键节点，在聚划算发展历程中审核共经历了以下三个主要演化阶段。

① 小二选品：买手根据自己的行业经验从报名商家和商品中挑选，整个质量把控和议价都是由小二完成，这导致商品质量参差不齐，小二工作量巨大，经验不可重用等问题。在 2012 年聚划算联合数据产品部门发起了商家和商品指标开发项目，随着商家指标、商品爆款分等指标上线，聚划算选品进入了一个新的时代。

② 小二数据化选品：2013 年，聚划算商家指标和商品爆款分等上线让小二从盲选的泥

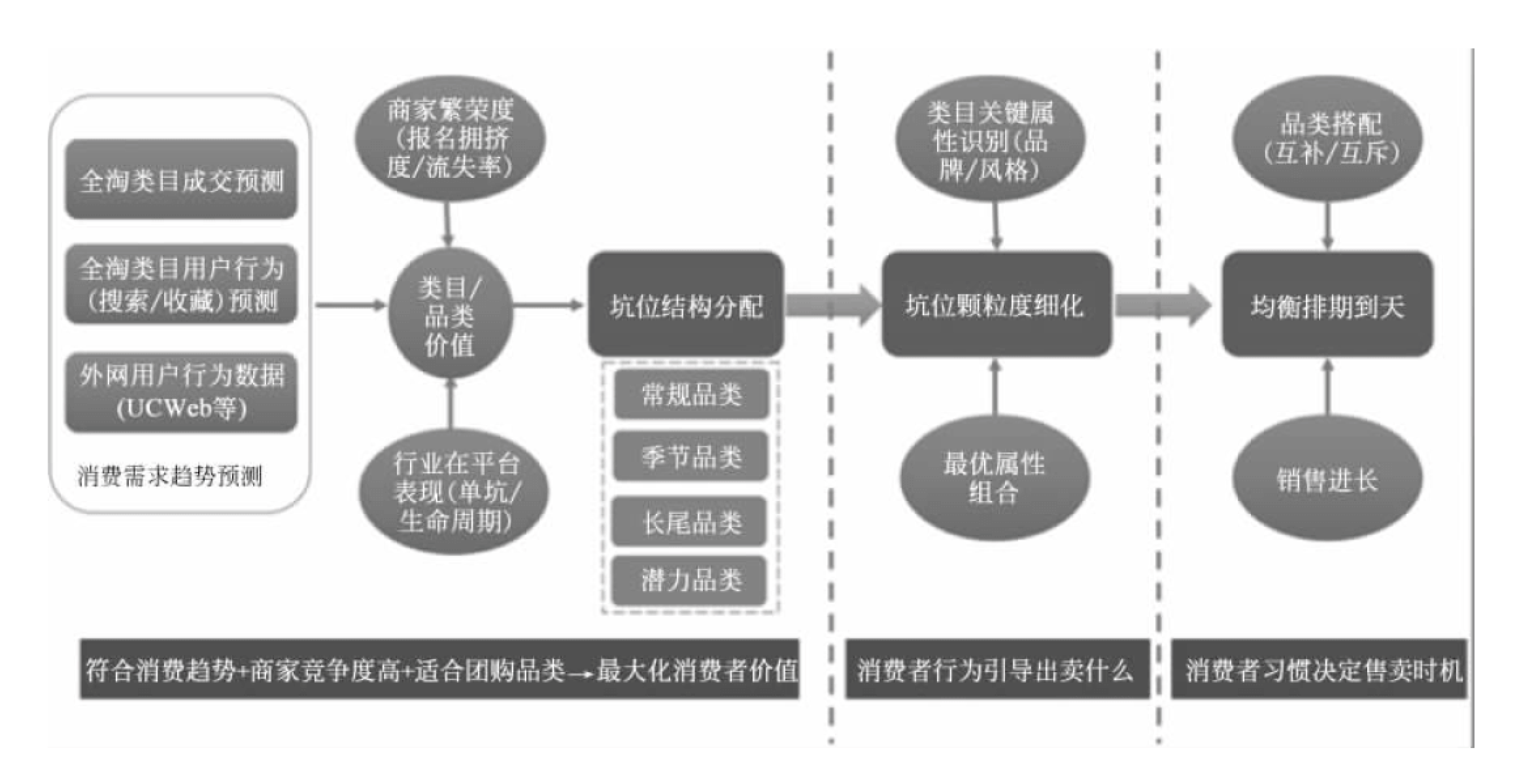


图 6-8 品类规划

潭中解脱出来。小二通过商家指标和商品爆款分等对报名商家和商品进行筛选，小二的行业经验又反馈到数据部门进一步优化算法模型。随着广泛使用，另外一个问题逐步凸显出来，统一的商家和商品指标准确率浮动很大，各个类目的个性化要求越来越多，商品爆款划分迎来了第二个转折点，算法模型按类目拆分，按照淘系一级类目拆分为 21 个模型，融入各类目各行业的特性，指标的准确率和稳定性大幅提升，一切都准备就绪了，可以开启新篇章了。

③ 数据自动化选品：2014 年是阿里的 DT 元年，也是聚划算数据化运营的元年，通过大数据来做出决策，让小二回归商业，成为行业营销专家。聚划算选品也做出了一个大胆的决定，通过数据自动化选品，面临的第一个问题是有商家指标和商品指标两个大指标，还有一些品牌指标等其他指标，数据决策必然只有一个决策标准，合并所有的指标，诞生了新的选品分指标，选品分数也采用了按行业类目特性来计算权重，但类目划分粒度更细、更加合理。对选品的离线评测机制也同步开展。为自动化跨出了重要的一步。面临的第二个问题，也是最难的问题：控价。淘系的商品价格，优惠非常复杂，而且商品同款非常多，要准确识别一个商品的历史最低价和全网最低价本身就是一个难题，自动做出合理的议价就更难了。针对此问题，在一淘产品库基础上识别商品近 90 天最低价，首先保证上聚商品的价格是自身的最低价。虽然实现了自动过滤价格，不过合理议价和保证全网最低价才是聚划算的最终目标。

聚划算业务流程如图 6-9 所示。

总体来看，数据化运营在自动化组织网站运营上取得了一定的效果。但在这个过程中，收获更多的是对于数据化运营的思考。先有网站自身的定位，才有数据化运营。数据化运营的核心，是找到一系列对数据获取、处理和利用的方式，直逼最终目标。网站需要有一定的机制设计来与各个角色进行数据交互，模型偏向于运营经验的沉淀，算法是利用这些经验和数据的手段。我们在看到数据化运营这几个字的时候，很容易把数据化运营放在运营人员和网站后台系统范围来看，而不是学会“怎样通过数据来看市场和用户的变化，怎样通过数据来影响或者改变，怎么获取更多更有用更真实的数据并形成”，数据化运营未来是设计

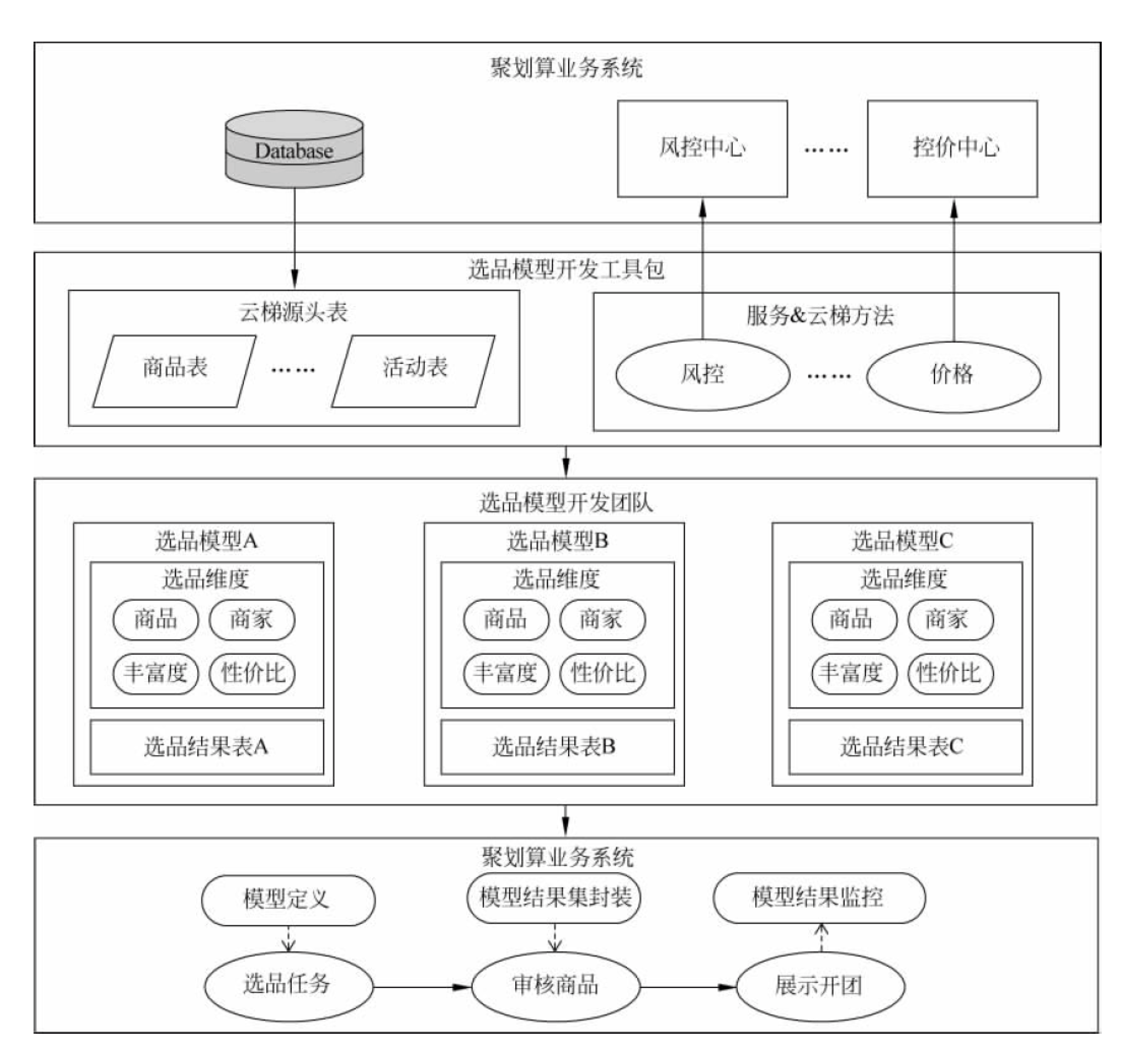


图 6-9 聚划算业务流程

一套有效的机制能获取并帮助平台以及各种角色利用数据。

6.2.2 消费者端：数据化导购

在消费者端，推荐“好宝贝”的方法主要是两个场景，基于搜索和非搜索推荐，前者是淘宝主搜的方式，后者是聚划算的重点。因为消费者登录无线端 3～5min 如果找不到喜欢的宝贝就离开了，所以个性化的首屏展示十分关键。聚划算致力于将最适合消费者的商品主动推荐过来，并提升大流量在商品列表、详情页的订单转化率，基于偏好推测你最想看到什么商品，以及商品与商品之间的关联推荐（啤酒与尿布相关性），在商品详情页、付款成功页“趁热打铁”推荐更多相关宝贝，促进“交叉销售”效果。在无线端，比较私密的场景下，根据用户喜好等个体因素主导，将数据理论与业务经验相结合，独创在线学习的“马虎算法”，使用个性化展示，将多个团队开发的在线学习、个性化等算法融合，不断优化不同场景下的数据推荐策略，成功将转化率大幅提升。在 PC 端，注重调性与氛围，采用类似天猫的“赛马”机制，根据宝贝的历史表现（点击率、销售额等多项复杂因素），即某一时段的销售情况，将评估潜力足的宝贝往前排放展示，同样将订单转化率提升了很大比例。

聚划算每天收集到很多有价值的数据，日成交 UV 和成交订单已经超过百万，超过千万的浏览 UV，用户的行为可以成为数据驱动开发很重要的切入点，在此分享一些有趣的数据发现。

1. 购买前的踌躇

博弈论中有这样一个观点，“消费者的购买冲动是随着对商品接触次数的增多而减弱的”。也就是说，如果能让用户一时冲动就买下商品，成交的可能性会大大增加。当用户对商品了解越多，接触越多，也许一些负面的东西就留下来了，便无购买意愿。当然，用户对商品的多次试探，其中的原因可能也很复杂，也许是预算不够，也许是本身就没有这个需要，那么如此一来他也可能不会加入“购物”的大军。提取某天的聚划算 PC 交易数据，并整合了当天的浏览数据，就成交前的用户浏览次数和决策时间进行分析，如图 6-10 所示。

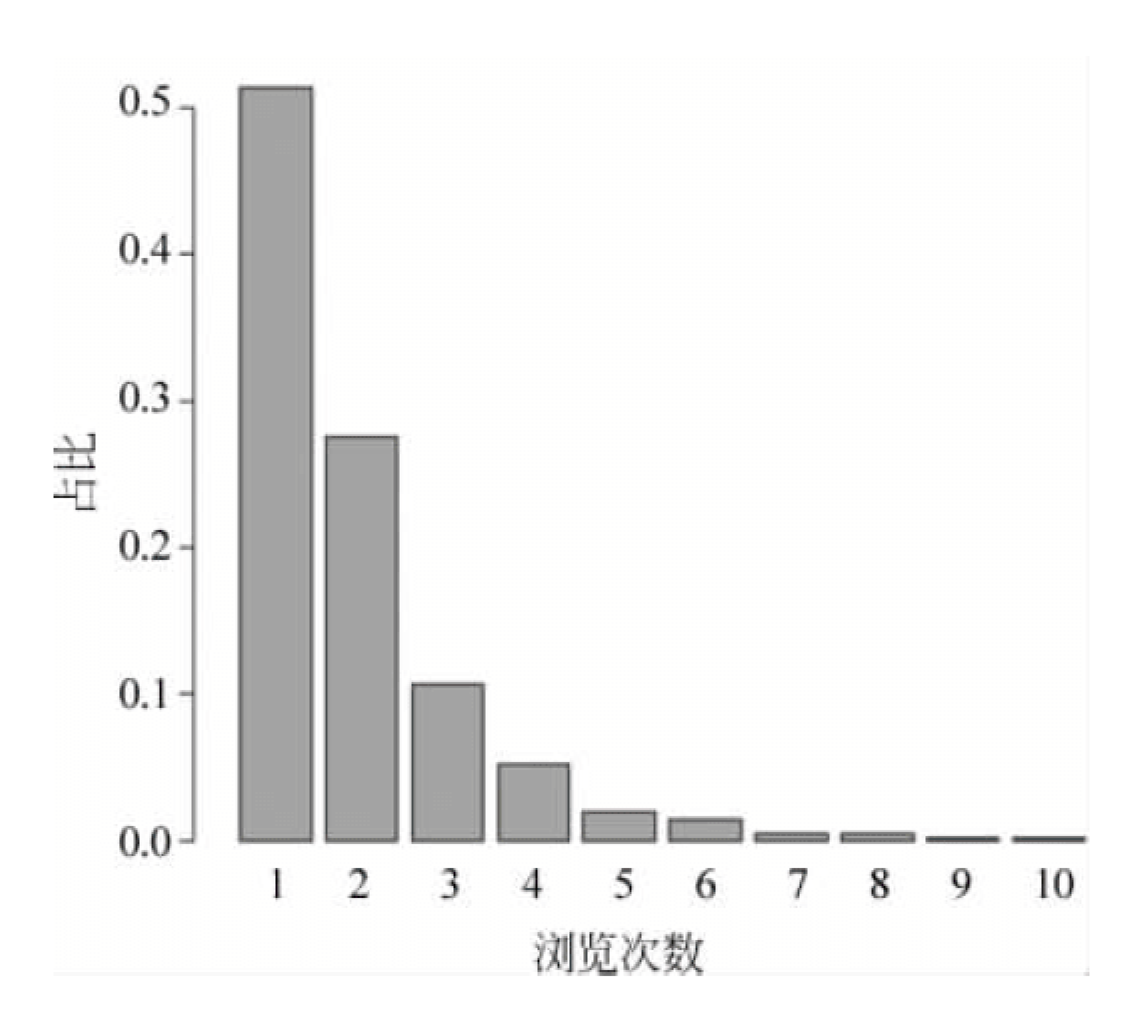


图 6-10　成交前商品详情页的浏览次数

从图 6-10 中可以看出，绝大部分用户是“冲动型”的。90%左右的用户，在三次浏览以内，就完成了下单和付款。浏览次数的加权平均值是 3.0，浏览三次以内成交的占比为 89.4%（占有浏览量的订单总数）。

图 6-11 是用户在下单付款前的“决策时间”分布图。这里所谓的决策时间，就是下单时间与当天第一次浏览的时间差。主流的详情页决策时长在 1～5min（50～300s）。时长的统计和浏览次数很不同，时长分布曲线是急速上升后缓慢下降。后面有很长的尾巴，说明决策时间在消费者中差异比较大，用平均值（平均值 1340.8s）作代表的意义就不大了。

在完成以上两个简单分析后，有一个问题一直困扰着数据分析师，在交易和日志记录中，每天存在着约 5 万～8 万没有任何浏览记录的订单（这个数字在查询日志扩展到成交前一天后依然没有太大的变化）。在数据团队讨论后，做出了两种猜想：一是有一些用户在手机端加入购物车，然后在 PC 端进行的付款，无法完整跟踪；二是有一批“专业”的刷单账号，通过聚划算的直接购买链接（buy_item_action）采用机器下单，批量作弊。于是数据专家对这些用户产生了兴趣。

2. 交易行为聚类

聚类方法有很多种，数据科学家采用了网络科学中的社团发现算法，并综合考虑了商品

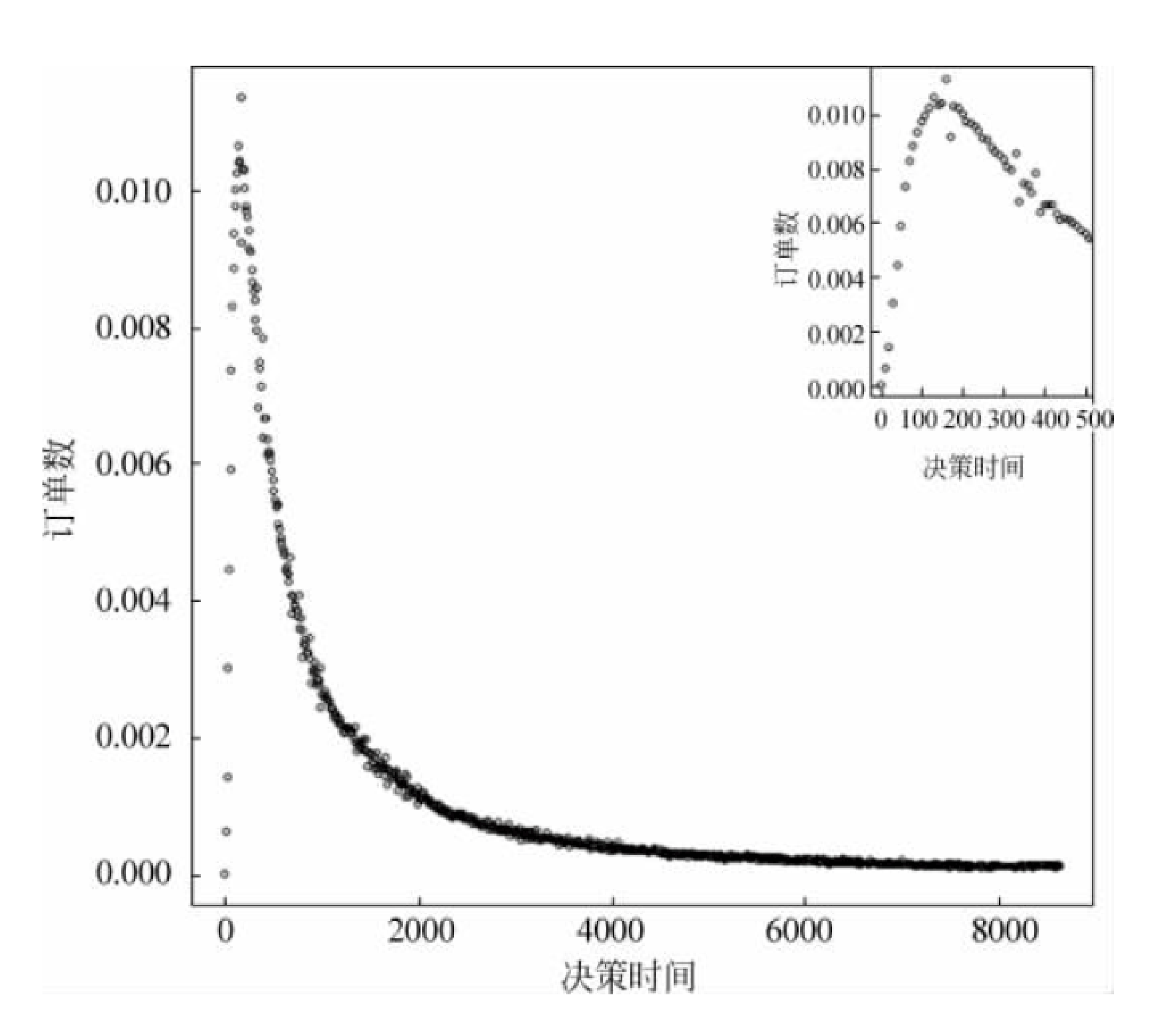


图 6-11　成交前决策时间分布

和时间的因素，对用户进行聚类。网络生成过程如图 6-12 所示。

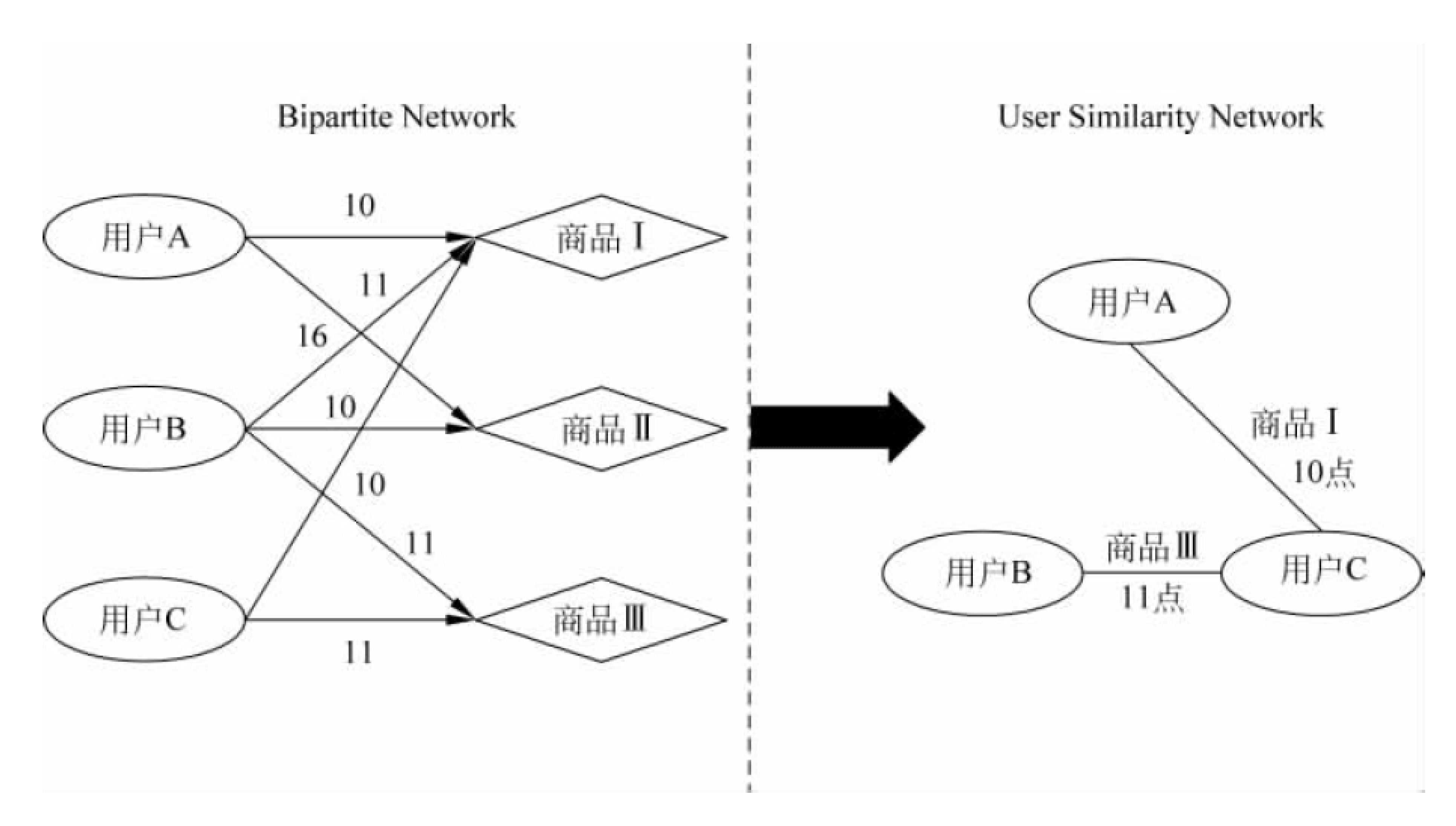


图 6-12　用户聚类网络生成过程

首先数据团队拿到的是用户的交易行为，即(用户-商品)的二分图。通过规则投影，形成(用户-用户)的网络。用户连边的规则是：如果两个用户，在相同的时间(按小时划分)购买了相同的商品，则建立一条边，如果重合次数不止一次，则边权表示行为重合的次数。于是抽取了某天无浏览记录的 56 467 个用户，并建立了他们的行为相似网络，如图 6-13 所示。

大部分社团是小的，只有少数是大社团。成员在 100 以上的社团仅有 15 个，成员有 10～100 个人的社团有 1192 个，那么依然有 8658 个社团是极小的社团。数据专家对大社团进行一些分析，首先是最大的社团，这个社团有 689 个用户，却只有 10 个商品(其中两个只出现一次交易，在图中被省略)。这个群体基本可以定性为青春女性。

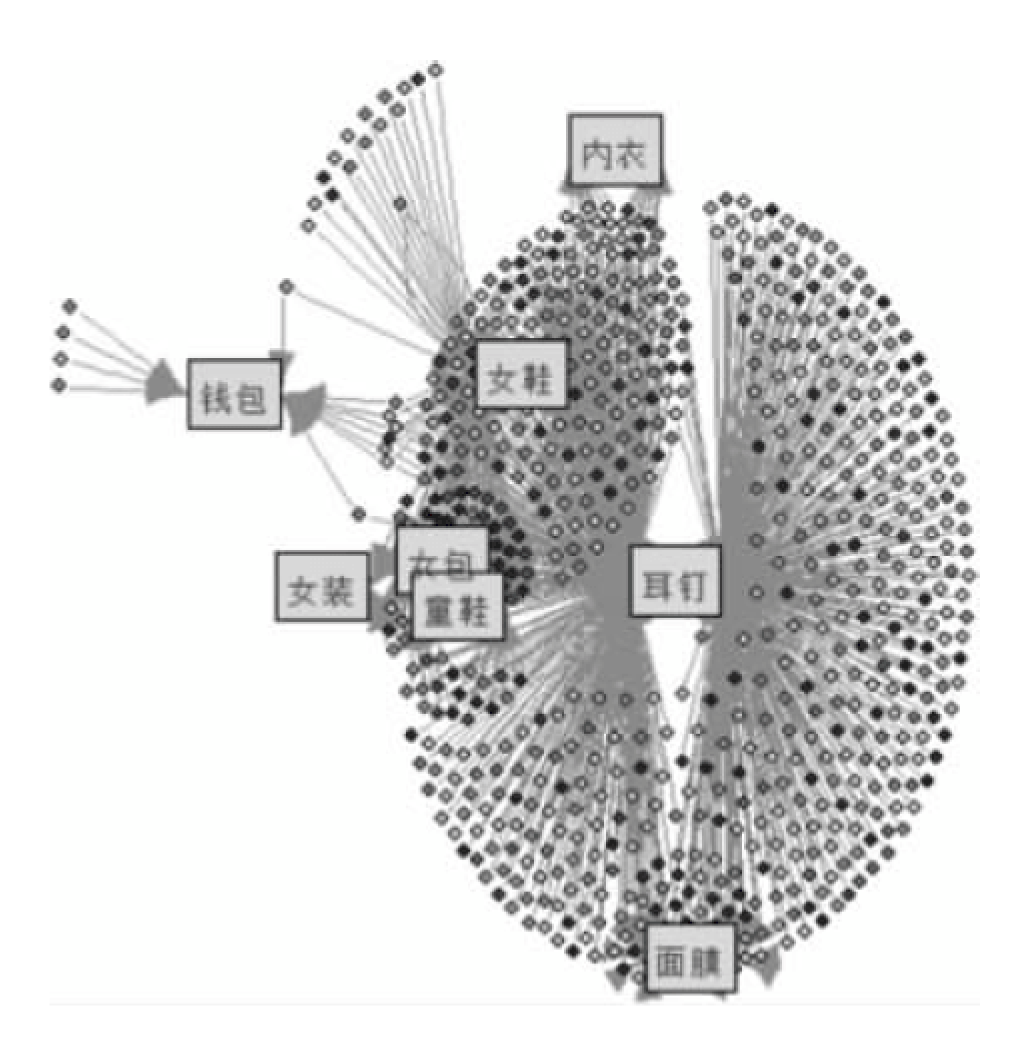


图 6-13　用户相似行为网络

对剩下的几个 TOP10 社团进行商品查询，得到的情况列在表 6-1 中。

表 6-1　用户社团 TOP10

社团大小	购买类别	简单定性
689	耳钉、护理	青春少女
642	女鞋、女包	年轻女性
624	女装、香水	年轻女性
563	耳钉、女装	年轻女性
514	睡衣、T 恤	家庭女性
481	睡衣、戒烟、养胃	中年女性
422	剃须刀、戒烟产品	家庭男性或主妇
199	面部护理套装	同店消费
175	女鞋、女裤	女性
150	面膜(御泥坊)	同店消费

除了根据兴趣划分外，我们还可以发现，在 100 大小的社团中，出现了几乎所有用户的交易都在同一个商店里面。不得不使我们猜想，有没有可能是商家做的鬼？雇用一批水军来刷销量。它提供给我们一个线索，就是可以通过商品和时间的维度来辅助虚假交易的监测和预警。同时，上面的社团分析，让我们猜测虚假交易的用户往往容易形成具有较强的关系网络(比如熟人、小号)，这在日后的数据产品中，被作为一个重要的参考来源。

3. 聚划算秒杀反作弊

聚划算的电商云上已经储存了丰富的业务数据，系统数据也因为集团的 DT 技术布局而唾手可得，多种数据的交叉复用，往往能带来很多意想不到的价值。通过尝试解决聚划算秒杀器作弊问题的契机，数据团队利用交易网络的聚类分析，尝试了多种数据跨平台的交叉复用，取得了不错的效果。

图 6-14 是交易棱镜系统的整体架构，云端(Yuntil＋ODPS)和线上应用通过 DataX 和 TT(TimeTunnel)的交互，实现了业务数据、系统数据交叉复用的闭环。云梯负责存储和离线计算，DataX 负责跨介质同步数据，TT 负责收集线上日志到云端。

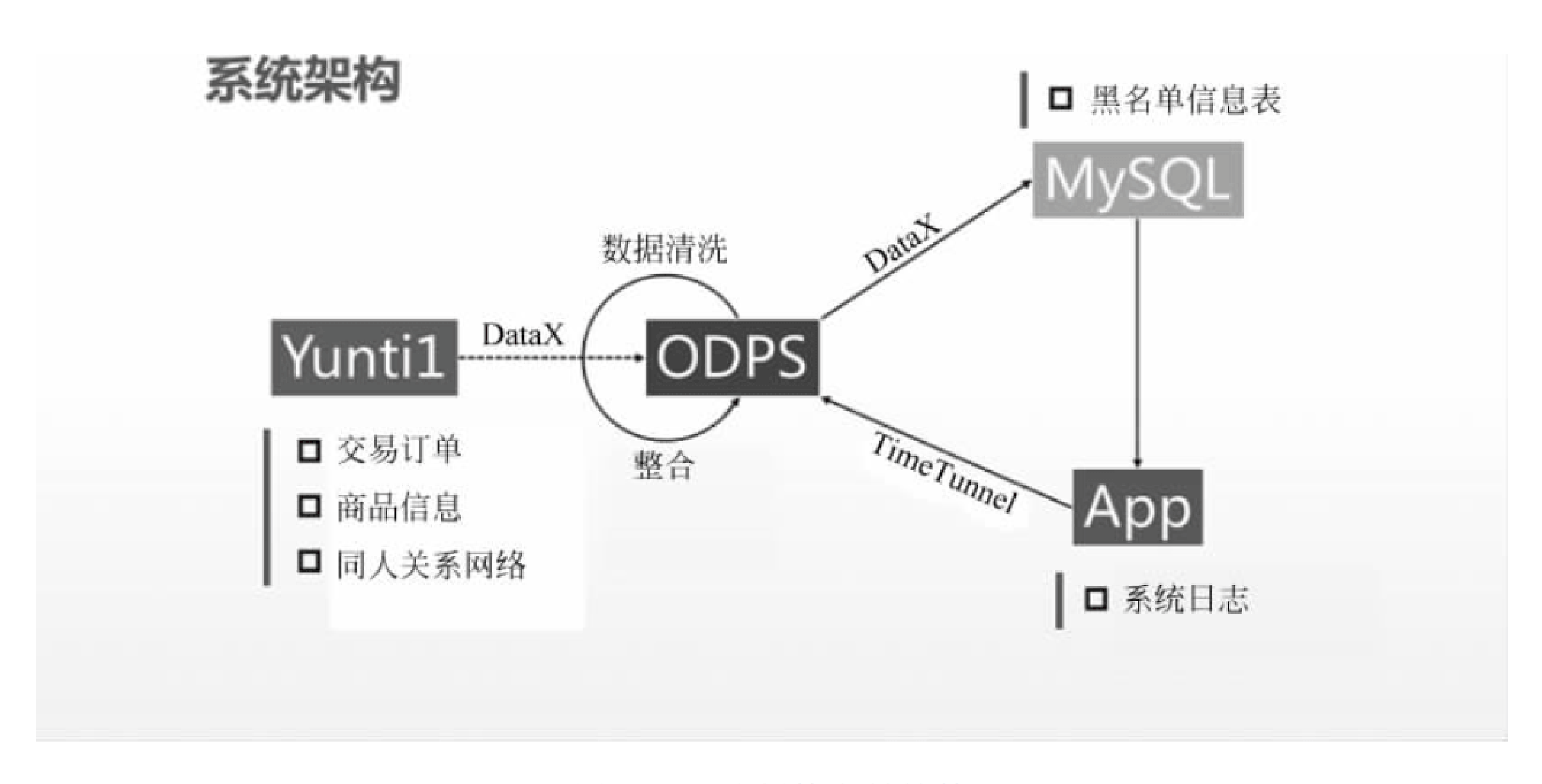


图 6-14　聚划算交易棱镜

ODPS 是数据的中心，通过整合来自云梯 1、ODPS 和线上返回的日志，计算出新的黑名单数据，再用 DataX 回流到 MySQL 数据库中，由线上应用进行调用。DataX 能够在不同介质进行数据搬运，包括云梯 1 到 ODPS、ODPS 到 MySQL 以及 TAIR。

所谓秒杀活动，原本是重在参与、生死由天的玩法。但是我们从数据分析一下，发现聚划算竟然存在这么多零秒订单，如图 6-15 所示。

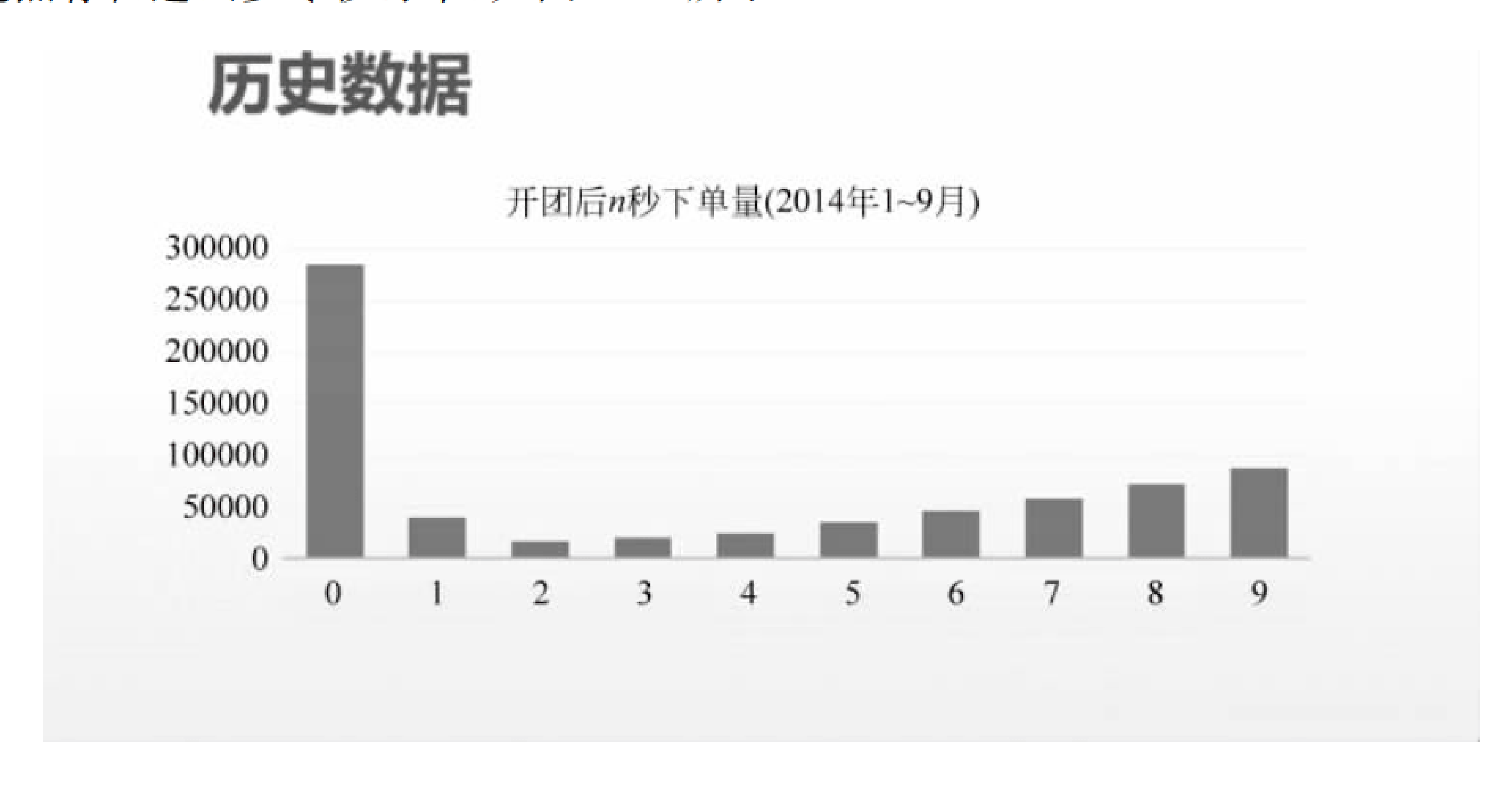


图 6-15　历史数据

由于历史原因，聚划算 PC 交易在单击“马上抢”按钮后，会跳转到天猫或者淘宝集市的宝贝详情页进行 SKU 选择和下单。普通用户手动的速度，几乎不可能在 1000ms(＝1s)以内完成从参团到下单至订单生成的全过程。而这些 0 秒订单，只有一个解释，就是利用了秒杀软件。最直截了当的方法，就是监控从聚划算参团到天猫或淘宝详情页的订单生成时间，如果是 1000ms 以内，就把这个订单 kill 掉。

如图 6-16 和图 6-17 所示分别为数据架构图和应用场景图。

我们通过引入秒杀恶意名单和同人关系网络，构造了一个专属聚划算交易的名单库，有效地维护了聚划算的交易环境，让更多普通用户有机会抢到心仪的商品。



图 6-16 数据架构图

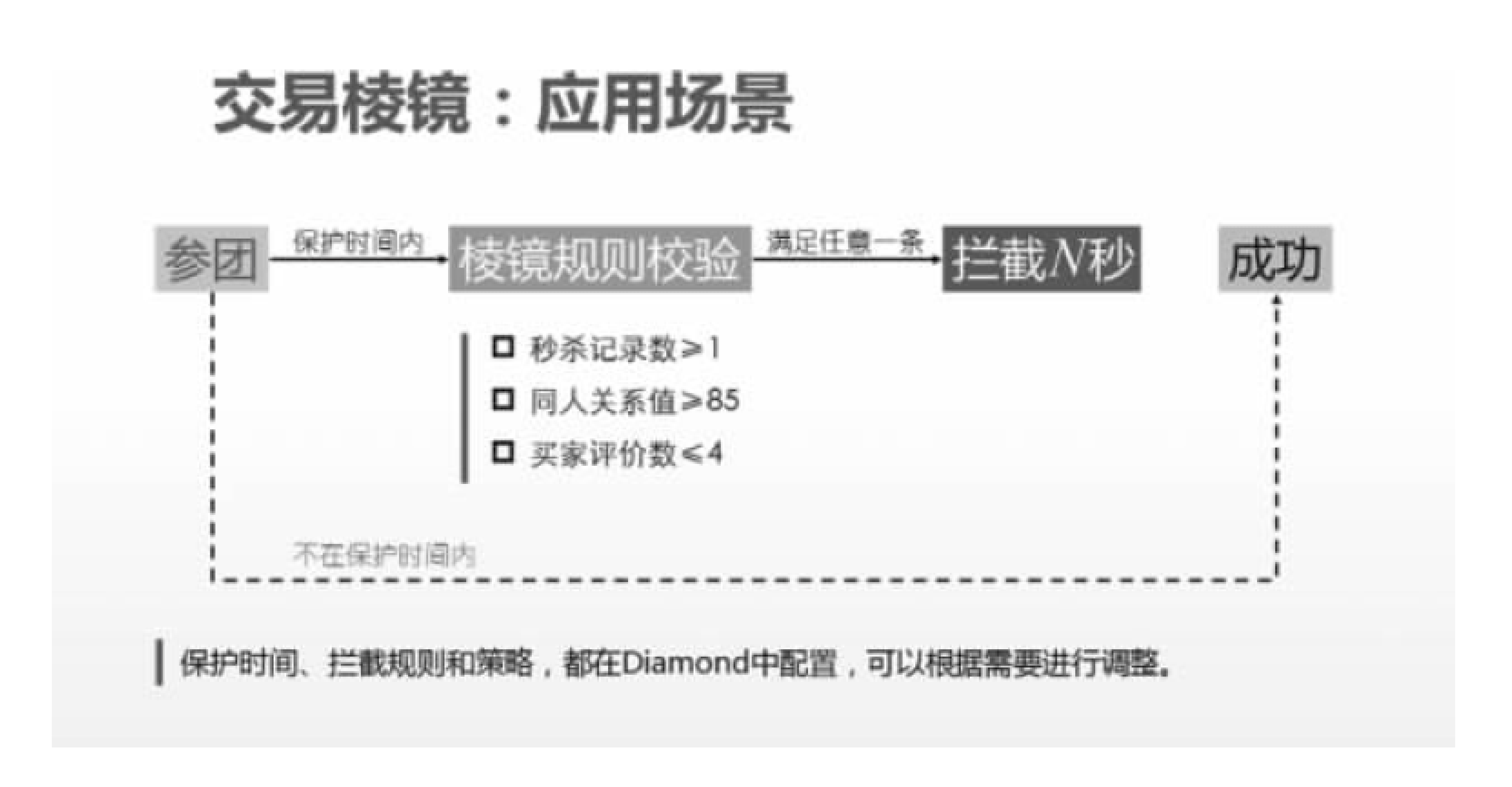


图 6-17 应用场景图

秒杀拦截系统第一天上线，拦截量就达到顶峰（大多数秒杀器毫无准备），如图 6-18 所示。随着作弊用户的察觉，每天的拦截量伴随着零秒订单数量逐步减少，现在每日的零秒订单维持在 1000 个以下。

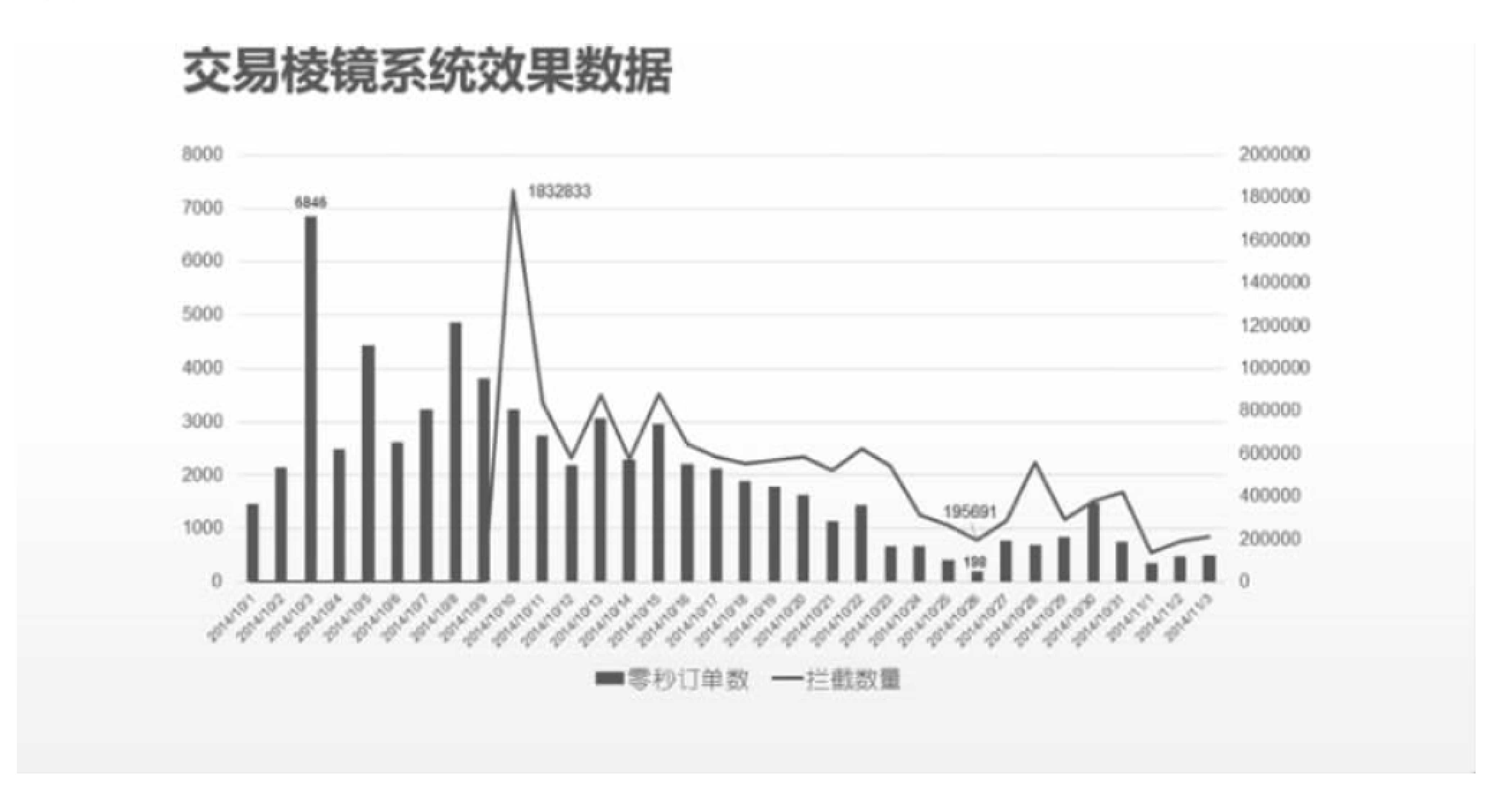


图 6-18 效果报告图

6.2.3　“聚划算”的数据价值

1. 业务价值

通过大数据算法设计与优化，将聚划算订单移动端转化率提升了 20%～30%，PC 端转化率提升了 10%以上，零秒订单减少了 65%以上，极大地提升了商品销售量，大幅降低了作弊比例。

2. 产业价值

为移动电商、在线团购营销，提出了数据营销、数据招商、数据反欺诈的创新模式，为 DT 时代的产业形成行业应用标杆，具有良好的商业复制价值。

3. 社会价值

保障消费者公平、便捷、舒适地享用电商团购服务，保护了优质商品和信誉良好商家的商业利益，培育更健康的电商市场秩序，DT 社会化应用功不可没。



第 7 章 让数据告诉你“谁可信”

虚拟的网络世界存在着秩序混乱和诚信缺失的危机，大数据技术的出现采集和存储了每个人、每个时点、每个位置、每个事件等信息，从而使虚拟网络的每个节点上的描述不再是单一独立的。三人成虎的故事不会再演，因为我们可以获得全城人的信息，甚至老虎自己都会告诉你“他没进城”。

“谁是可信的?”这个问题，大数据可以给出回答。

7.1 “区块”成“链”

1982 年，一位美国计算机科学家莱斯利·兰伯特（Leslie Lamport）给大家讲了一个故事。

拜占庭位于如今的土耳其的伊斯坦布尔，是东罗马帝国的首都。由于当时拜占庭罗马帝国国土辽阔，为了防御，每个军队都分隔很远，将军与将军之间只能靠信差传消息。在战争的时候，拜占庭军队内所有将军和副官必须达成共识，决定是否有赢的机会才去攻打敌人的阵营。但是，在军队内有可能存有叛徒和敌军的间谍，左右将军们的决定又扰乱整体军队的秩序。在进行共识时，结果并不代表大多数人的意见。这时，在已知有成员谋反的情况下，其余忠诚的将军在不受叛徒的影响下如何达成一致的协议，这就是著名的“拜占庭将军问题”。[①]

在原始的战争中仅能采用“出行靠走，通信靠吼”的口头或信件传递，所以拜占庭将军问题的含义是：在存在消息丢失的不可靠信道上试图通过消息传递的方式达到一致性是不可能的。

随着统计、计算等技术的出现和发展，困扰拜占庭将军的问题已不再是问题。2008 年 11 月 1 日，比特币之父中本聪发表了一篇题为“Bitcoin P2P e-cash paper”（比特币：一种点对点的电子现金系统）的文章，阐述了基于 P2P 网络技术、加密技术、时间戳技术、区块链技术等的电子现金系统的构架理念，这标志着比特币的诞生。两个月后理论步入实践，2009 年 1 月 3 日第一个序号为 0 的比特币创世区块诞生。几天后 2009 年 1 月 9 日出现序号为 1 的区块并与序号为 0 的创世区块相连接形成了链，标志着区块链的诞生。

区块链是一个分布式账本，一种通过去中心化、去信任的方式集体维护一个可靠数据库的技术方案。在区块链的思想中，通过“工作量证明链”就可以获得解决“拜占庭将军问题”的方案。

① 维基百科.

7.1.1　区块的形成

从数据的角度来看，区块链是一种几乎不可能被更改的分布式数据库。这里的“分布式”不仅体现为数据的分布式存储，也体现为数据的分布式记录（即由系统参与者共同维护）。从技术的角度来看，区块链并不是一种单一的技术，而是多种技术整合的结果。这些技术以新的结构组合在一起，形成了一种新的数据记录、存储和表达的方式，如图 7-1 所示。

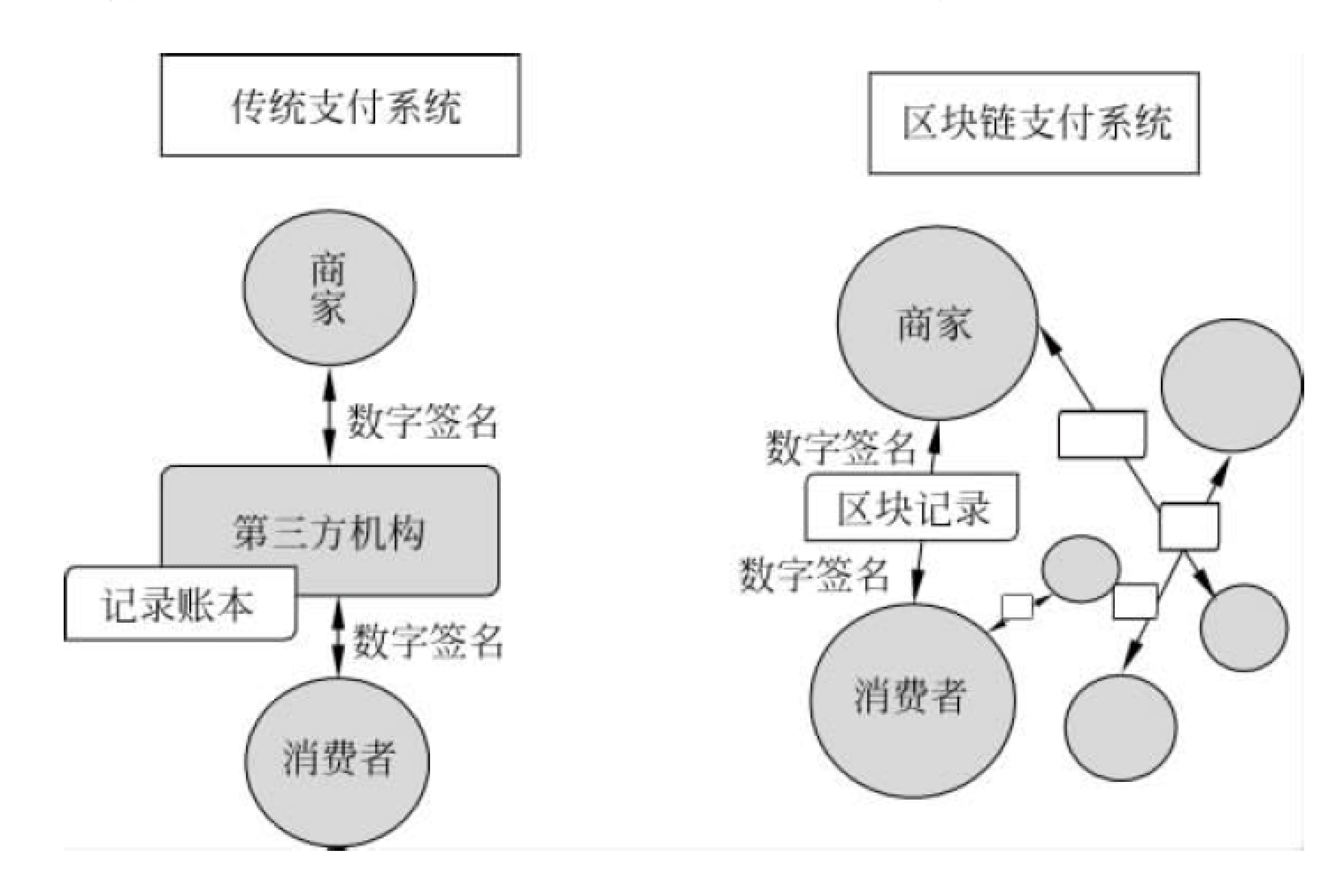


图 7-1　区块链

1. 区块

在区块链技术中，数据以电子记录的形式被永久储存下来，存放这些电子记录的文件就称为“区块（Block）”。区块是按时间顺序一个一个先后生成的，每一个区块记录下它在被创建期间发生的所有价值交换活动，所有区块汇总起来形成一个记录合集，这个合集就是区块链。其基本思想可以这样理解：通过建立一个互联网上的公共账本，由网络中所有参与的用户共同在账本上记账与核账，每个人（计算机）都有个一样的账本，所有的数据都是公开透明的，并不需要一个中心服务器作为信任中介，在技术层面就能保证信息的真实性、不可篡改性，也就是可信性。

某个区块好比账本中的一页，对于普通账本，信息被写在了纸上，而对于区块链，数据能够通过保存于区块中被永久地记录在数字货币网络上。区块上的数据，一旦书写就很难被修改或移除。区块的结构包括区块大小、区块头、交易计数器和交易，如图 7-2 所示。

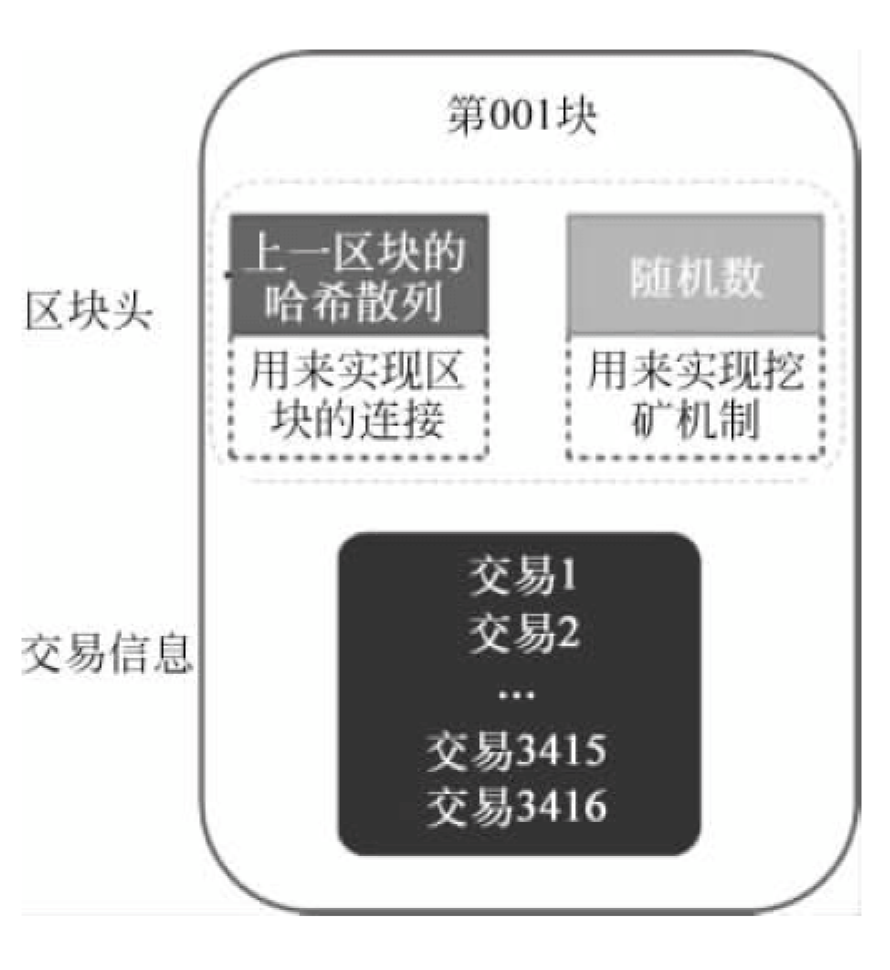


图 7-2　区块

区块大小，顾名思义，所表示的是其所在区块的大小，通常为 4B。

区块头，包含除了交易相关信息以外的所有信息。包括引用父区块哈希值（复杂的随机值）的数据，用于将该区块与区块链中前一区块相连接；记录该区块产生的近似时间的时间戳；用来有效地总结区块中所有交易的 Merkle 树根；用于跟踪软件或协议

更新的版本号；与区块工作量证明算法相关的难度目标和计算器。

交易计数器记录了该区块中包含的交易的数量。

交易记录了该区块中交易的信息。

2. 区块"链"

区块链以区块为单位组织数据，如图 7-3 所示。全网所有的交易记录都以交易单的形式存储在全网唯一的区块链中。

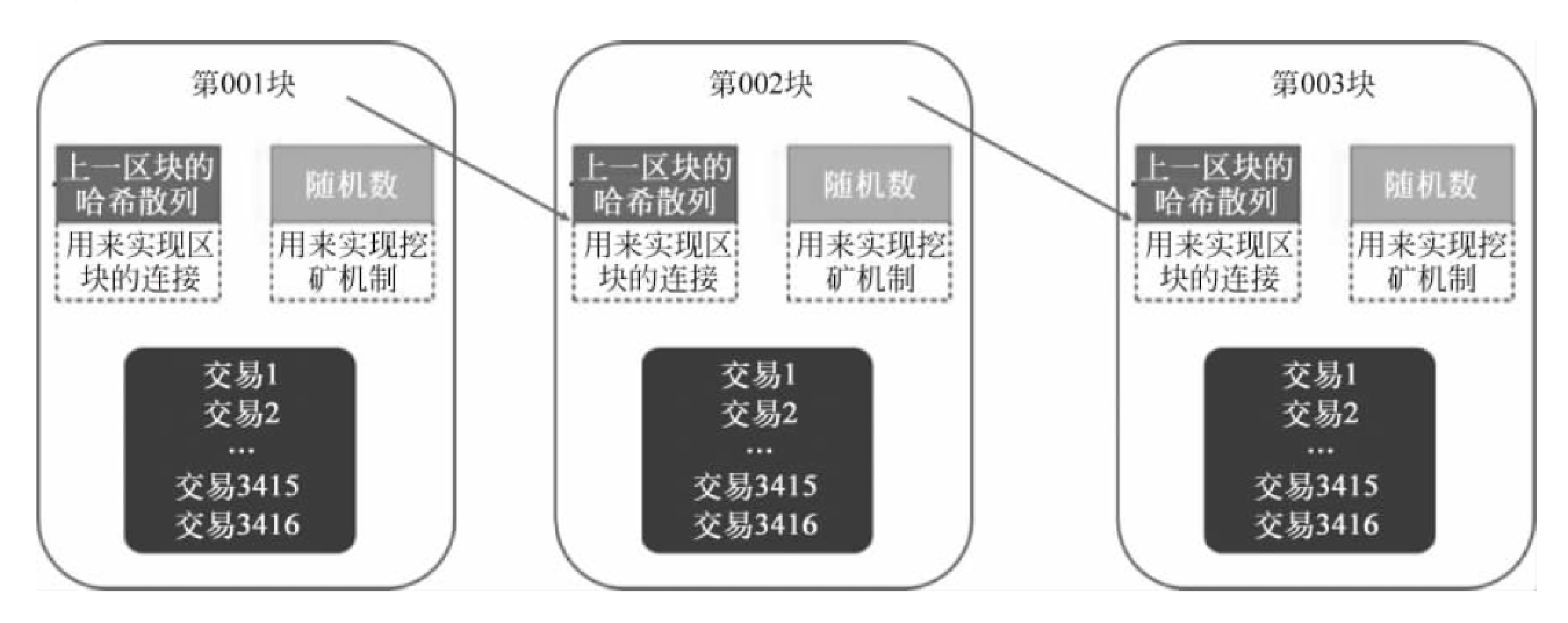


图 7-3　区块"链"

7.1.2　区块链的特征

区块链的特点及发展来源于它所产生的土壤——互联网技术的发展和云计算、大数据的兴起。

1. 去中心化

区块链系统是由大量节点共同组成的一个点对点网络，不存在中心化的硬件或管理机构，任一节点的权利和义务都是均等的，系统中的数据块由整个系统中所有具有维护功能的节点共同维护，且任一节点的损坏或者失去都不会影响整个系统的运作。

2. 共识信任机制

区块链技术从根本上改变了中心化的信用创造方式，运用一套基于共识的数学算法，在机器之间建立"信任"网络，从而通过技术背书而非中心化信用机构来进行信用创造。借助区块链的算法证明机制，参与整个系统中的每个节点之间进行数据交换无须建立信任过程。在系统指定的规则范围和时间范围内，节点之间不能也无法欺骗其他节点，即少量节点无法完成造假。

3. 信息不可篡改

区块链系统将通过分布式数据库的形式，让每个参与节点都能获得一份完整数据库的备份。一旦信息经过验证添加到区块链上，就会永久地存储起来，除非能够同时控制整个系统中超过 51%的节点，否则单个节点上对数据库的修改是无效的，因此区块链的数据可靠性很高，且参与系统中的节点越多，计算能力越强，该系统中的数据安全性越高。

4. 开放性

区块链系统是开放的，除了交易各方的私有信息被加密外，区块链的数据对所有人公开，

任何人都可以通过公开的接口查询区块链数据和开发相关应用，因此整个系统信息高度透明。

5. 匿名性

由于节点间无须互相信任，因此节点间无须公开身份，系统中每个参与的节点都是匿名的。参与交易的双方通过地址传递信息，即便获取了全部的区块信息也无法知道参与交易的双方到底是谁，只有掌握了私钥的人才能开启自己的“钱包”。此外，在诸如比特币的交易中，提倡为每一笔交易申请不同的地址，从而进一步保障了交易方的隐私。

6. 跨平台

区块链网络上的节点是基于共同的算法和数据结构独立运行的，主要消耗的是计算资源，与平台无关，可以在任意平台部署计算节点。

区块链是一种伴随着比特币出现的思想，比特币只是区块链的第一个应用。有了区块链之后，当一个用户想要进行历史交易的验证时，可以通过一系列基于密码学与数据结构学的运算追踪交易所属的区块，从而完成验证。

7.2　“芝麻信用”让信用等于财富

区块链体现的是群体智慧，是互联网思维的技术实现。大数据则提供了群体智慧的来源，两者的结合可以实现一种群体评价体系。简单地说，大数据时代要证明一个人的诚信或一个事实的真伪，需要所有人共同的认定。

当被问到“如何证明我爸是我爸”的时候，不是由公证处、警察局、人事户籍单位来开具证明文件。而是通过我的邻居、家里的七大姑八大姨、我爸单位的同事、小学班主任是否看到我爸去开家长会等共同来论证的。

7.2.1　什么是信用

大数据在金融领域中最重要的应用或者说是颠覆性的作用，就是可以构建新的去中心的信用评价体系。

什么是信用？信用是经济活动的金融借贷，随着受信人的信用值从低到高发生变化，信用值很低的人交易时必须“一手交钱、一手交货”，信用值高的人则可以享受“先交货，再付款”等高可信服务，如图 7-4 所示。信用极大地提升了消费环境的便利性，作为一种基础，有助于促进诚信社会的建立。



图 7-4　信用高低的区别

欧美和日本等国征信行业发展历程表明经济发展与信用体系相辅相成的密切关系，伴随我国经济快速发展，征信行业将进入快车道。按照宏源证券分析师的预测，中国个人征信市场空间为 1030 亿元，而目前个人征信和企业征信的总规模才 20 亿元。

英美征信行业经历了一百多年的发展，已形成稳定的市场格局和完备的法律体系，个人征信以 Equifax、Transunion、Experian 三大征信公司为主体，企业征信以邓白氏公司为主，如图 7-5 所示。



图 7-5　英美征信产业

(1) 数据提供方：个人、机构自愿提供基础数据，例如金融机构、电信运营商、法院、公共事业单位等机构，另外，一千多个地方信用局（三大信用局的外包机构）也在不断收集信用数据。

(2) 信用评估方：以“三大”为首的信用局，加工数据（付账记录、未偿还债务、开立账户时长、贷款、使用过的信贷记录），依托 FICO 模型来赋予个人或企业“信用分数”。

(3) 信用服务使用方：在这一环节信用产生价值，不同组织有偿使用信用服务，例如在金融服务中，开新账户、银行卡申请、房车贷款、保险等领域广泛应用，其他工作生活领域也是以信用为基础的，包括公共事业、电话安装、就业升职、房车租赁、工商注册。

与美国在 1860 年成立第一家信用局相比，中国在 2004 年才迟迟启动个人征信系统建设，2006 年 11 月人民银行征信中心在上海正式注册为事业法人单位，截至 2014 年年底，央行个人征信系统共收录 8 亿多用户（其中有信用记录的约为三亿）。

从国外征信业历史来看，我国征信业起步晚。2015 年 1 月 5 日，中国人民银行在《关于做好个人征信业务准备工作的通知》中，要求芝麻信用管理有限公司（蚂蚁金服旗下）等 8 家机构做好个人征信业务的准备工作，这些公司在日常经营过程中积累下来的自有商业性数据，可通过购买、交换、自采以及免费获取等方式获得信用信息和数据，前提是收集和分析这些数据需要获得个人授权，遵守法律规定，民营征信企业的加入让征信业真正焕发了市场化和大数据的活力。

7.2.2　从“信用”到“财富”

马云说：“阿里巴巴真是希望让信用等于财富。几年前，我们呼吁银行全力支持中小企业，但是银行有自己的难处，它们的模式很难让它们真正地去服务好网商、服务好中小企业，所以阿里用互联网的思想和互联网的技术去支撑整个社会未来金融体系的重建。在阿里的金融体系里面，我们不需要抵押，我们需要信用；我们不需要关系，我们需要信用；我们不需要你挣多少钱，我们需要你踏踏实实地为客户服务。”

芝麻信用正通过信用数据平台以及各种场景中的“信用支付”服务产品构建“信用金融”体系，通过互联网金融帮助普通消费者、中小企业更便捷地获得更强的经济活力与更佳的体验，重构涵盖个人信用和企业信用的生态系统有着重要意义，具有互联网基因的企业在这方

面优势巨大。

如图7-6所示，“芝麻信用”致力于为13亿中国人、6000万企业法人建立信用档案，整合网上银行、电商、社交、招聘、婚介、公积金社保、交通运输、搜索引擎，最终聚合形成个人身份认证、工作及教育背景认证等维度的信息。传统金融征信只是瞄准信贷业务，而“芝麻信用”对接神州租车采用个人信用背书，让高信用消费者（芝麻分高于650）“无抵押”快捷租车，对接“去啊”在旅游中帮助讲信用消费者（芝麻分高于600）“零押金”享受“信用住”。



图7-6　蚂蚁金服的征信基础

1. 信用评分

2015年1月28日，蚂蚁金服旗下的个人征信机构芝麻信用首推“芝麻信用分”（简称“芝麻分”），这是中国有史以来首个个人信用评分，可以直观呈现用户的信用水平，如图7-7所示。

在“支付宝钱包”中打开“财富”栏便能看到芝麻信用分的选项，芝麻信用采用了国际上通行的信用分来直观地表现信用水平，如图7-8所示。

芝麻信用分最低350分，最高950分，分数越高代表信用程度越好，违约可能性越低。这与美国的FICO分评分（300～850分之间）非常相近。

如何计算？芝麻信用分的计算基础主要包含用户信用历史、行为偏好、履约能力、身份特质、人脉关系5个维度，如图7-9所示。

（1）信用历史：过往信用账户还款记录及信用账户历史。

（2）行为偏好：在购物、缴费、转账、理财等活动中的偏好及稳定性。

（3）履约能力：享用各类信用服务并确保及时履约。

（4）身份特质：在使用相关服务过程中留下的足够丰富和可靠的个人基本信息。



图7-7　支付宝钱包中的芝麻信用分

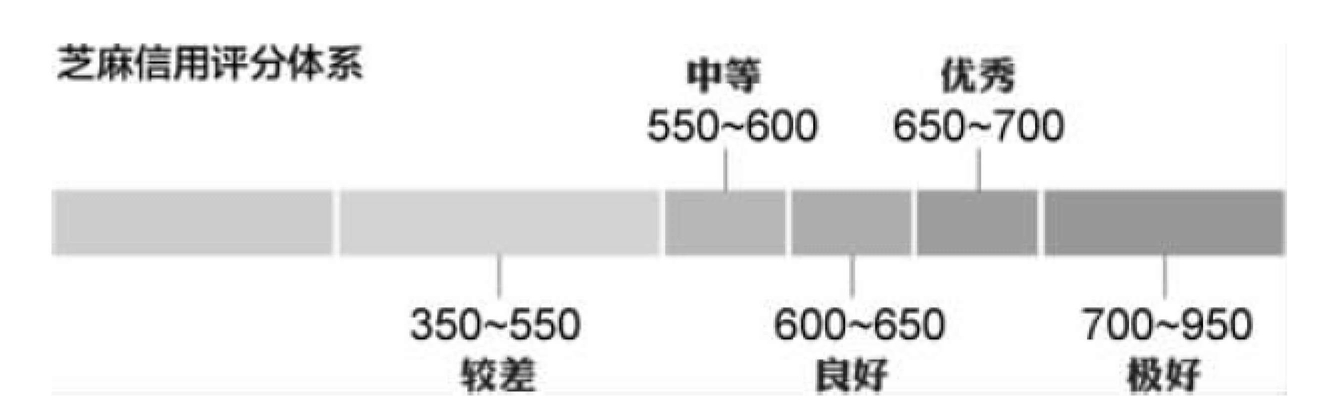


图 7-8　芝麻信用评分体系



图 7-9　芝麻信用分评估维度

(5) 人脉关系：好友的身份特征以及跟好友互动程度。

数据来源上，除了会接入阿里巴巴集团的电商数据和蚂蚁金融服务集团的互联网金融数据外，芝麻信用还与公安部（身份证）、教育部（学历学籍）、工商总局（企业注册）、质检总局（组织机构代码）等众多公共机构以及合作伙伴建立了数据合作关系，同时也将开辟各类渠道允许用户主动提交各类信用相关信息，如图 7-10 和图 7-11 所示。

芝麻信用数据涵盖了信用卡还款、网购、转账、理财、水电煤缴费、租房信息、住址搬迁历史、社交关系等方方面面。人们在日常生活中点点滴滴的行为，都与信用息息相关，信用开始影响人们的生活服务模式。

芝麻信用认为，信用是整个社会的基础设施，芝麻信用会被应用到生活的方方面面，而不仅局限于金融领域，如图 7-12 所示。出国签证时使馆根据信用分快速签批，无须再准备资产证明、收入证明；旅游旺季根据信用分，住酒店前无须预付“担保留房”，住酒店免押金

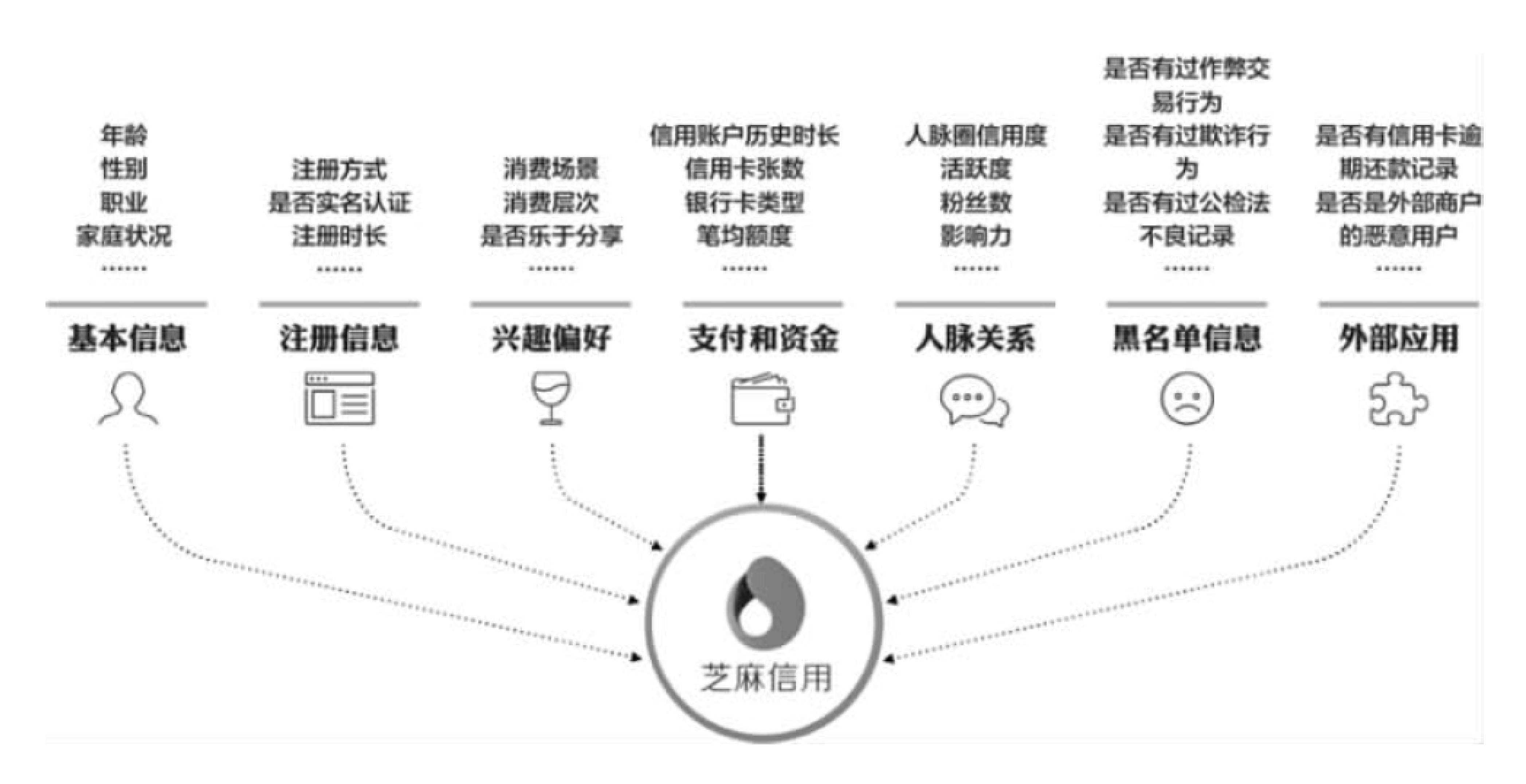


图 7-10　芝麻信用的数据内容(示例)

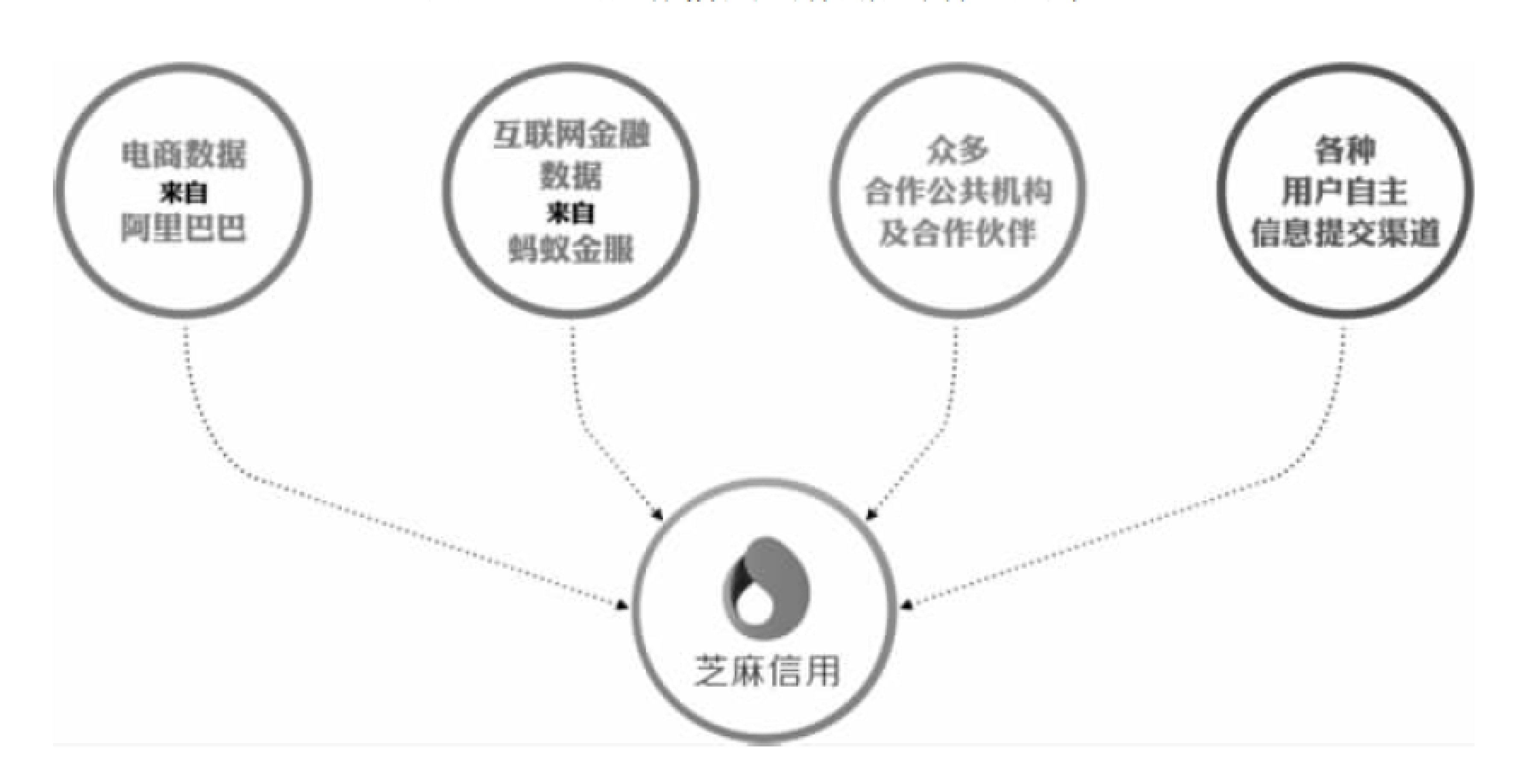


图 7-11　芝麻信用的数据来源



图 7-12　芝麻信用生活服务

入驻；租车时根据信用分取消“押金”及“预授权”环节；未来，讲信用的人会享受到更丰富的省时、省心、省钱的社会服务，例如招聘时是否录用、贷款是否发放、奖学金是否给予、约会交友是否继续等都可以先查一查对方的芝麻分，如图 7-13 所示。



图 7-13　信用就是财富

目前，芝麻信用已经跟租车、租房、婚恋、签证等多个领域的合作伙伴进行合作，对外提供出行、住宿、金融、购物、社交、民生等多种信用相关的便捷服务。这意味着当你的芝麻分达到一定数值，租车、住酒店时可以不用再交押金，网购时可以先试后买，在各国大使馆办理签证时不用再闪转腾挪办存款证明，贷款时可以更快得到批复、拿到比别人低的利率，甚至相亲时也可以最大程度避免婚骗。但如果借钱不还、恶意购物，这些行为也可能引发找不到工作、租不到房、申请不到贷款、找不到女朋友等连锁反应。

2. 信用“财富”的困惑

（1）会不会一失足成千古恨？

芝麻信用的答案是：正常情况下，一个人的信用信息是相对稳定的。当然，如果关键信用发生重大变化，这种变化也会很快反映到芝麻分上。不过，信用的提升是一个循序渐进的过程，无法通过单个行为或事件迅速提升，需要长期积累。

芝麻分是芝麻信用根据当前了解的信息，运用大数据方法综合评估而得，个人用户通过让芝麻信用了解其更多信息将有助于提升芝麻分。同时在日常生活或经济活动中尽可能使用信用服务及时履行约定也有助于芝麻分的提升。

（2）我的信用信息会不会人尽皆知？

芝麻信用负责人表示，不管是机构还是他人，要查看你的芝麻分，都必须获得你本人的授权。此外，芝麻分是通过对海量信用信息的综合评估和处理，得出一个信用分数，去直观地呈现信用水平。这种呈现形式，可以更好地保护个人的具体信用信息和隐私。

截至 2013 年年底，央行银行个人征信系统中收录有信贷记录的自然人约 3.2 亿，历经 16 年从无到有、细致全面地构筑了中国企业、个人的征信数据库，并与所有商业银行、农信社联网运行，是所有工商业正常运营的信任基础。

芝麻信用等民营征信机构所覆盖的网民草根群体，是对传统征信机构有益的市场数据补充，如未有过借贷、未申请过信用卡的人，学生群体、蓝领工人、个体户、自由职业者等。他们以这些人群的网络行为轨迹数据为基础，利用大数据技术和数据分析模型评估出其信用等级，与现有征信系统形成了很好的互补效应。

7.2.3 信用商圈

芝麻信用可以应用在哪些领域?

芝麻信用不仅在金融借贷关系中有很强的信用风险预测能力,而且在生活服务中也具备很好的区分能力。因此未来在金融领域,需要押金、预授权的租房、租车、酒店等行业,新兴的分享经济领域,婚恋、交友等生活场景等领域都可以用到芝麻信用,如图7-14所示。芝麻信用今后还能够产生服务于不同商业领域的信用报告。



图7-14 未来的信用商业生态

如图7-15所示为市场化征信机构给社会带来的改变。



图7-15 市场化征信机构带给社会的改变

1. 商业价值

中小企业征信是个难题,尤其在国内,人行的征信中心只能通过银行获取企业的贷款和还贷信息,对于很难从银行获得融资的中小企业,人行掌握的信息几乎真空,阿里巴巴具备中小企业互联网交易信息,在该领域具备十几年的商业数据积累。

2. 社会价值

美国的“FICO”是全球最著名的信用打分私企。大部分美国人都会有一个属于自己的分数,高于一定标准,申请信用卡或办贷款就会一路顺风;低于一定标准,相同的申请就可能困难重重,需要比分数高的人多提供一大堆材料。欧美信用局参考日常生活的很多方面数据,例如购物出行、电费水费甚至地铁逃票都会成为影响信用的潜在因素。很多时候我们

之所以感觉外国人更守规矩，并不是全部因为他们道德崇高，而是整个信用环境都在鼓励大家守规矩。中国的征信市场当然也能达到“FICO”这样的效果。除了央行征信系统之外，随着阿里、腾讯这些国内最有创造力的互联网企业加入，个人信用会成为越来越重要的个人标签。今后的人们不会直到信用卡逾期影响房贷才发现征信体系的存在，而是从平时生活中就开始有意识地为自己积攒信用，良好的信用又会反过来带给大家更多方便。传统信用机构与新兴民营征信机构的开放合作、社会协同会带给每个企业、百姓更好的社会公共服务与商业价值，真正在工作生活的每个领域做到“让信用等于财富”。



大数据“地图”

随着移动互联网时代的到来，越来越多的公司正在创造一种可能性，把虚拟网络世界中的大数据和地理信息位置结合起来。通过利用连接移动设备诸如智能手机、室内场地Wi-Fi网络、低成本的蓝牙通信功能以及其他几种特殊的技术，位置分析厂商已经使人们有可能获得位置分析解决方案，并能够快速获取信息，以很低的成本获取分析结果——追踪到客户，并把位置发送到供应商那里进行分析，通过一系列精密的仪表，获得可操作的数据访问，最终实现精准营销策略。

数据“地图”不仅是地理空间的地图，而且涵盖着地图上每个点的服务信息和每时每刻，处在不同位置上的每个人的需求和偏好信息，因此数据地图既是个人生活出行的便捷工具，也是智慧城市建设的好帮手。

8.1 便捷交通大数据服务

当前，大数据已经上升为国家战略，为此国家2015年发布了一系列的指导性文件。2015年6月，国办发〔2015〕51号：《国务院办公厅关于运用大数据加强对市场主体服务和监管的若干意见》；2015年8月，国务院关于印发《促进大数据发展行动纲要》的通知；2015年年底，国务院常务会议通过了《关于促进大数据发展的行动纲要》，强调将大数据打造成新常态下经济提质增效升级的新引擎，为经济发展和社会进步提供更加有力的支撑。为此各政府部门、行业及相关企业陆续掀起了数据挖掘、分析的热潮。

北京市政交通一卡通有限公司经过多年的发展及运营，自2006年5月，其一卡通卡发卡量已超过9000万张，在公交、地铁、出租、加油站、公租自行车、停车场、公园、学籍、小额消费等十大领域24个行业得到广泛应用，系统日均处理交易约1600万笔，历史累积数据440亿笔，交易详细记录了出行OD(交通起止点)、出行时间、交通工具之间的换乘及线路等信息。结合大数据理念，开展对海量业务数据进行深加工，挖掘公共交通和商业运行数据信息，对政府部门、公众和商家提供信息参考和咨询服务变得尤为重要。表现在交通、城市规划等政府部门对于市民公共交通轨迹分析及出行规律需要一卡通数据做支持；一卡通公司内部对于分析业务实质，优化业务流程，需要大数据分析为决策提供数据支持；对于运营业主和商户，在决策和商户精准营销方面也迫切需要提供数据支持；而结合持卡人的使用习惯和需求，通过数据分析为持卡人提供更好的服务体验和更加多样化的产品形态，提供方便快捷的换乘路线，提升用户体验和黏性。在此背景下，北京市政交通一卡通有限公司自2014年基于互联网、大数据和云计算等技术开始探索对此海量数据的开发和利用，并产生

了一系列的成果。

8.1.1 城市公共交通存在的问题及其现状

随着城市化的不断发展，城市人口也快速增长，从而导致城市出行人口的大幅上升，进而引发“乘车难”等社会问题。造成这些问题的原因是多方面的，其中最根本的原因就是城市公共交通体系的不完善。

面对这些问题，政府及有关部门也出台并实施了一系列的解决措施，但“乘车难”等问题还是城市的一大难题。

就目前来看，城市交通是阻碍城市发展的重要因素之一。加大力度关注与发展城市公共交通，是解决城市交通问题的重要方式，政府方面也明确提出优先发展城市公共交通的发展策略。但是由于我国城市公共交通起步较晚，投资较大，范围较大等影响因素，使我国城市公共交通仍属于探索发展阶段，仍存在一些问题。

公共交通是城市的基础建设行业之一，它不仅具有生产性的特点，同时还具有服务性及公益性的特点。城市公共交通是影响整个城市社会经济发展的基本因素，但随着经济体制改革深化发展以及社会主义经济体制的确立，公交管理体制中的各种问题不断暴露出来。具体来说，主要有以下几个方面。

首先，在特定时间段、特定地点，堵车成为城市居民心中之痛。目前我国城市公共交通形式比较单一，主要是以公交车为主。公交车的运营密切关系着大部分城市居民的出行。但是在特定时间段，例如上下班高峰期；特定地点，如中小学门口、购物商场门口等都成为长时间、大规模堵车的重要时间点、地点。和国外交通发达的国家相比，我国的城市公共交通面临着更艰巨的任务。众所周知，中国是世界上人口最多的国家，相对于城市实际人口的快速增长，城市公共交通的缓慢发展还存在着差距。

其次，公共交通部分路线的规划安排不甚合理，一些路线公交车过于密集，而另一些路线公交车又较少，甚至缺失。根据有关调查，许多城市普遍存在公交车线路设置重复，一条路线上有近十几条公交车运行，而城市边缘居民地区、街道尚未开通公共交通路线，使这些地区居民日常出行非常困难。

最后，公共交通管理不到位，缺乏科学合理的管理。虽然社会各个方面均认识到公共交通是解决城市交通问题的重要途径之一，但是在实际的实施过程中，却没有完全严格执行。在城市公共交通规划中，公共交通整体规划不到位，管理跟不上，缺乏科学的管理方式、专业的管理人员。

8.1.2 大数据服务应用

1. 应用方式

公共交通大数据服务从城市一卡通清算中心交易库、客户数据库、公交行业数据库、轨道交通行业数据库、消费数据库、停车行业数据库、互联网业务数据库等抽取数据，对这些结构化和非结构化的数据进行抽取、清洗、整合、转换，存入共享数据库。

统计分析通过OBIEE(Oracle Business Intelligence Enterprise Edition)采用一定的算法和模型等读取处理数据，结果保存在服务层数据库，为用户提供可信的数据，还可通过可视化以各种统计图展现出来，通过PC、手机可以看到结果。

系统提供可视化的操作界面，用户可自己定义统计和参数，系统计算分析后给出对应的图表。

2. 技术原理

数据分析平台采用领先的 Hadoop 架构，有效结合 NoSQL、关系型和列式数据库的特性，同时有效预留基于内存技术的新一代数据库。

ETL 抽取工具根据一卡通数据特征研发，数据分析工具根据需求分期研发，使用 Echart 实现各种统计图表的生成，数据展示平台采用 Java 开发，B/S 模式。

3. 平台功能

数据分析平台提供了完整的统计分析功能，通过分类法、回归分析法、关系规则法、Web 数据挖掘法等，包括来自于统计学、机器学习、人工智能等方面的分析算法和数据模型，包括如关联、分类、预测等完整的全面挖掘分析功能。能够按照时间段、卡号段进行统计，能够根据卡号进行轨迹追踪，能够从复杂的数据集合中发现新的关联规则，继而进行深度挖掘，得到有效用的新信息。平台功能包括数据 ETL 模块、数据处理模块、数据模型、数据报表和数据展示 5 大部分功能，如图 8-1 和图 8-2 所示。



图 8-1　平台数据展示部分界面截图

4. 数据抽取

ETL 的目的是形成一个具有统一视图的干净的数据库，这个数据库包含分析所需要的所有数据。

ETL 的过程可以归纳为以下几个阶段。

(1) ETL 抽取策略。包含 ETL 架构、关键 ETL 过程中采用的方法、关键组件的定义、数据质量管理策略、项目流程等。平台中 ETL 抽取策略采用增量抽取和全量抽取相结合的方式，对于数据量较大且更新频率高的表，在抽取时采用增量抽取方式，如一卡通的交易流水信息；对于数据较小，更新频率不高的表，抽取时采用全量抽取，如 POS 机信息、行业代



图 8-2　平台数据展示部分界面截图

码信息等。

（2）ETL 抽取的 job（作业）。在数据整合设计的过程中，所有的设计人员将遵循数据整合策略设计中约定的方法。数据整合策略设计过程是一个高度迭代的过程，需要设计人员不断地根据一卡通数据分析的具体数据状况调整和优化设计，而且在设计完成后，需要 ETL 开发人员不断地根据业务数据的实际情况反馈修改意见进行修改。

（3）ETL 的 job 开发。这个开发过程需要不断地用实际的业务数据验证是否设计已经充分考虑了所有的业务流程。

（4）集成测试、数据质量测试。集成测试是为了保证大量的 ETL job 可以彼此协同完成数据整合过程，这个阶段更多的是对 ETL 调度和 job 之间的接口的测试；数据质量测试是数据整合的核心，这个阶段的数据质量测试是整个项目的数据质量管理的一部分，数据质量测试和其他数据质量管理过程中的阶段一样，需要业务人员和 IT 人员的密切配合，而且需要测试人员深刻理解数据质量管理过程中的最佳实践。

5. 数据清洗转换

此阶段主要是根据具体系统数据需求确定清洗原则，主要包括：数据检查与稽核；数据类型统一转换；空数据赋默认值；数据排序与拆分；脏数据处理等。

通过对 ETL 过来的源数据进行一系列处理生成符合数据仓库结构的数据，再导入数据仓库中。

6. 数据加载

根据业务分析的需求，通过构建数据模型，将清洗的数据按照数据模型进行装载。

便捷交通的核心是人，让公众便利获得多维度、准确便捷的出行信息服务是便捷交通的一个重要出发点，当前市场的常用服务仅提供路径选择和粗略的旅程时长估算。北京市政交通一卡通公司正在探索利用历年累积的海量数据，挖掘基于出行路径、出发时间点、旅程时长、拥挤度、价格等维度的历史出行规律，并结合天气状况、车辆限行、是否工作日等因素，为公众提供个性化的最佳出行匹配方案。该方案可允许使用者输入“最快捷”“最便宜”“最舒

适(拥挤度)”“最环保”等单一参数或复合参数,依据不同的出发时点提供不同的公共交通(含公租自行车)出行路线规划,以满足市民对出行越来越高的便捷、高效、绿色、舒适等需求。

8.1.3　个人应用场景

上述出行方案与优网科技的实时手机信令解决方案相结合,将产生更佳的聚合效应。出行中,利用手机的定位功能,在远端通过云计算监控出行路径中的拥堵情况与预测结果的拟合度,实时计算途经路段相应的交通状况,对预计换乘的节点进行预警,通过 App 实时优化既定出行方案,为持卡人提前进行拥堵疏导建议,使公共交通出行更为便捷、灵活,如图 8-3 所示。

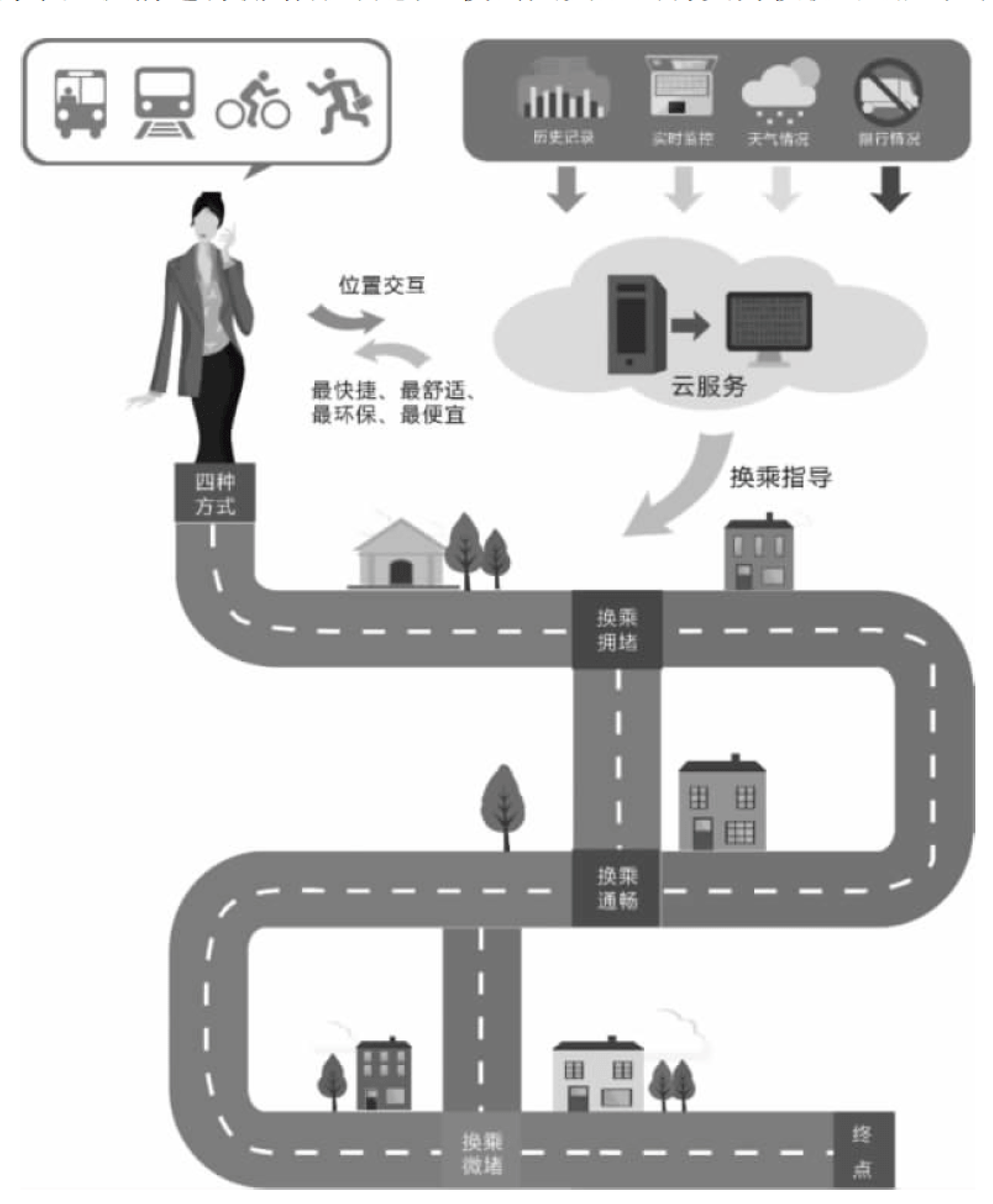


图 8-3　公共交通出行 App 场景描述

出行前:某天,晴,星期一,预计上午 9 点从 A 地出发,乘坐公共交通工具去 B 地,打开 App 线路查询功能,输入“最低价”为第一选择、“最快捷”为第二选择、“最舒适(拥挤度)”为第三选择。App 调用一卡通数据分析平台,利用十多年公众实际出行记录搭建的出行模型,筛选条件为“星期一”“天气晴朗”“上午 9 点”从 A 地出行至 B 地的全部样本,为其迅速推荐最佳出行线路规划,如图 8-4 所示。

出行中:利用手机信令数据,App 实时计算未完成行程的公共交通状况、未换乘节点的拥堵状况,并根据各路况现状随时优化出行方案,如图 8-5 所示。



图 8-4 公共交通出行 App 线路推荐



图 8-5 公共交通出行 App 线路优化

从我国城市实际情况来看，优先发展城市公共交通是正确的战略思想。全面建设一个高效的城市公共交通是城市发展的方向，如何应对公共交通发展中的问题，需要规划、设计、建设、运营、管理和服务等方面全面统筹解决，实施公共交通的可持续发展。一个城市只有具备了优质快速的公共交通网络，才会有一个通畅便捷的城市交通系统，才能为市民提供一个环保、安全、快捷、舒适、多层次的公共交通服务，才能真正促进城市的全面发展，才能提高城市的社会竞争能力。

北京作为特大型城市和首都，公共交通出行的规模、交通工具的类型和数量、拥堵的时间分布和程度在全国范围内都十分突出，问题解决的复杂程度也远超大部分城市。因此，北京市政交通一卡通公司在北京复杂的交通环境下，探索并部分实际应用的“互联网＋便捷交通”系列案例和方法论对其他城市具有借鉴价值，有利于“互联网＋便捷交通”在全国的落地和应用推广，增强全国公共交通领域的治理能力和提升公共交通运输的服务品质。

8.2 人群流动监控

2014 年 12 月 31 日上海外滩发生拥挤踩踏事件，造成 26 人死亡，49 人受伤的严重后果。据事后调查发现，当晚自 20 时期，外滩景区出现进多出少的情况，大量市民涌向外滩观景平台，呈现人员逐步聚集态势。20 时至 21 时约 12 万人，21 时至 22 时约 16 万人，22 时至 23 时约 24 万人，23 时至事件发生时约 31 万人。根据上海市政府新闻办公室发布的实时公共交通信息称截至 22 点 40 分，上海全路网客流已超过 1003 万人次，“再创历史新高”。

根据事后对事件的分析，发现导致事故的主要原因有：对新年倒计时活动变更风险未做评估；新年倒计时活动变更信息宣传严重不到位；预防准备严重缺失；对监测人员流量变化情况未及时研判、预警，未发布提示信息；应对处置失当。事后反思中也提到需要对事件监测预警，进一步提升突发事件的防范能力。

为避免此类事件的再次发生，建立人群流动趋势的实时信息系统成为一种重要需求。由于随着居民生活水平的提高和技术的发展，基本实现了人手一部移动终端的情形，除了传统的摄像头的画面监控的途径外，更可以通过移动终端信号、基站数据等途径获取到确定位置的人群聚集情况，通过建立数据分析平台，可以有效地监测在特定区域的人群聚集情况，有助于及时预警及提前准备好相应的应对措施。

基于对移动通信网络信令的数据分析，研发出人群区域分析系统，可以实现对密集区域的人流实时监控、预防告警等，指导其合理有效地了解人流流动趋势，避免事故发生。提供历史数据的环比、同比等不同的分析方法和展现，实现对人群流动的预测分析，实现有效的事前预警，事中处理及事后总结功能。

通过整合分析政府数据、用户的移动/固定网络数据、公共部门现有的各类数据，采用大数据分析技术，可以建立一套基于人群聚集和流动的信息化的监控预警系统，对城市的主要场所人流密集程度进行动态的监控，做到事前预警、事中处置、事后分析。同时通过不断积累数据、优化模型，逐步建立综合化的大型的监控、预警平台，对建设、交通、旅游和商业等多个领域部署监视，促进智慧城市、智慧交通，城市科技进步与产业发展。

基于大数据的人群流动监测系统整体构架如图 8-6 所示。

该人群流动监测系统的数据来源主要是用户位置数据、业务量数据、基站数据和用户互联网行为相关数据，并将已有的城市视频监控系统数据接入基于大数据的人群流动信息分

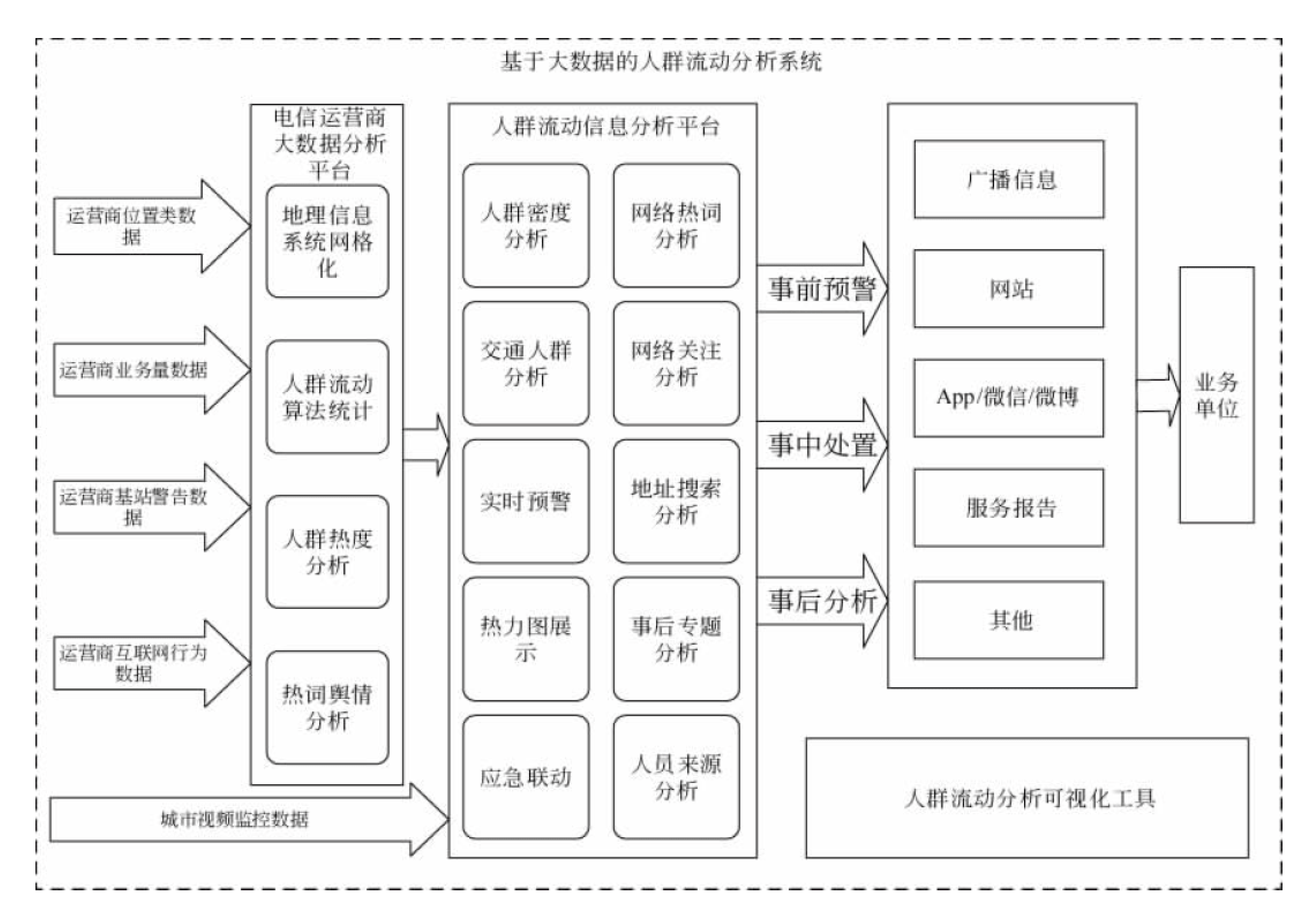


图 8-6　人群流动监测系统

析平台。通过相应算法可以获得相关区域的人群分布热度情况、区域用户数、网络关注热词分析、人员数据分析，并可以通过可视化模块实时动态地对用户分布数据进行展现。通过扩展相应的定制预警功能，通过多种形式提供相应区域内的信息预警服务，总体上可实现事前预警，按照区域灵活配置人群密度预警阈值、人群密集预警、通过历史数据对人流趋势智能预测；事中应急，通过信息发布平台提供相应信息发布、疏导用户、应急联动各相关部门；事后分析，事后结合记录的数据进行回放和实践总结。

利用新建预警平台回顾上海外滩事件的时候发现，从当晚 23 点人员分布的热力图看，外滩、南京东路站附近、浦东滨江大道、陆家嘴附近、天潼路附近是人流高度集中区域，其次是外滩源、南京路步行街、梅龙镇广场和八佰伴附近人流也较为密集。对比在陈毅广场、外滩源、陆家嘴三个主要基站的当天小时人数和历史平均人数水平发现，三个基站均在当天 19 点到 20 点，大幅超过历史平均人数的一倍以上，22 点到 23 点基站大数据达到峰值，超过历史平均人数的八九倍，陈毅广场和外滩源在 0 点达到峰值。从人员构成来看，本市人员的占比总数约为 40%，大部分来自外省如安徽、江苏、河南、江西、浙江等省份。人群流动统计如图 8-7 所示。

基于大数据的人群流动分析系统的建立，依托于通信企业所拥有的数据优势和技术优势，通过大数据的方法实现人群流动的监测、预警监测，布置在热门景区、人流量集中的区域，可以做到事前预警、事中处置和事后分析，有效地预防和降低类似安全事件的发生。

在这个案例中，需要在网上获得海量的用户信息进行计算并存储统计得到的历史信息。数据分布广泛、数据体量大、数据实时性与计算速度的要求很高，这些要求只有在大数据技

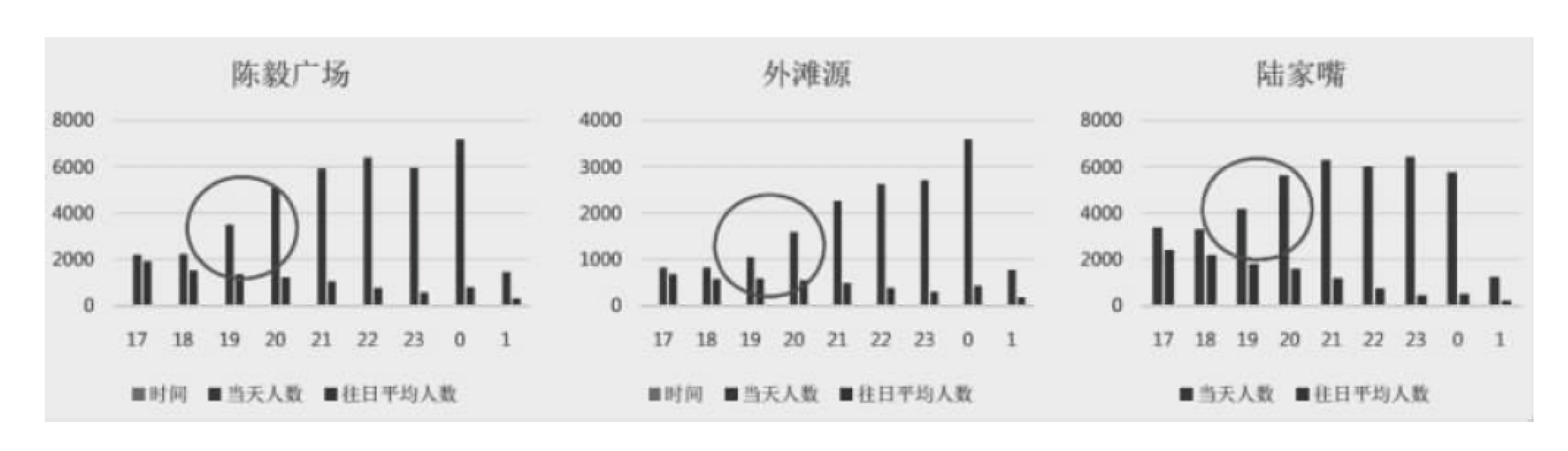


图 8-7　人群流动统计

术的帮助下才能完成。

8.3　实时车流控制系统

随着我国人民生活水平的不断提高，我国的汽车保有量在近几年一直处于持续上升的趋势，截至 2015 年年底我国的机动车保有量已达到 2.79 亿。交通拥堵越来越成为城市交通的一个难题。城市交通拥堵问题不仅为居民生活造成经济和时间上的双重负面影响，对于社会经济的发展也造成了巨大的损失，数据显示，因交通问题引起的社会经济的损失额约占 GDP 的 5%～8%。城市拥堵的问题已不再限于一线城市，已经逐渐蔓延到二线城市，而随着我国汽车销量和保有量的持续增加，国内主要的一二线城市将步入整体的交通拥堵时期。

随着机动车数量的增加，对交通和环境都带来了巨大的压力，由于汽车在拥堵情况下废气排放更加严重，治理拥堵对于减少汽车污染也很重要。汽车作为能源消耗和废气排放的大户，如何更好地做到节能减排已成为令人关注的问题。从近几年来雾霾问题较为严重的北京市来看，2015 年北京市的汽车拥有量为 561 万辆，排放污染物 70 万吨，而作为引起北京雾霾问题的一个重要测量因子 pm2.5 的来源数据看，31.1%的排放源来自于机动车。减少尾气污染除了在技术层面通过采用更先进的技术降低污染的排放外，通过更高效的汽车导航和交通信息服务也能够有效地减少汽车排放，从另一个方面缓解汽车带来的污染问题。

面对上述交通和环境问题，国内某通信企业基于自身拥有广泛的基站和海量数据的优势，结合自有硬件和技术优势，构建基于海量数据的大数据交通信息监测平台，通过数据共享等方式服务于政府有关部门、汽车厂家、地图导航企业等机构和商家，合作发挥数据的价值，共同为缓解交通和环境问题做出一定的贡献。

基于大数据的交通信息监测平台的总体结构框架图如图 8-8 所示。

该交通信息监测平台主要数据来源如下。

(1) 用户数据，特别是主动定位模块获取到的用户位置数据。

(2) 定位平台数据，包括基站数据、Wi-Fi 数据、固定电话分布数据及宽带接入网络位置数据等。

(3) 其他数据，来源包含政府、合作伙伴等，如地图导航终端的数据共享。

交通信息监测平台通过对获取的实时位置信息进行计算、分析、挖掘和存储，形成实时动态的交通信息数据流，通过标准化、可视化的输出模式为相应的客户提供交通信息、安全、终端数据等多维度的信息服务。同时对数据进行存储，结合历史数据对照分析做出相应的

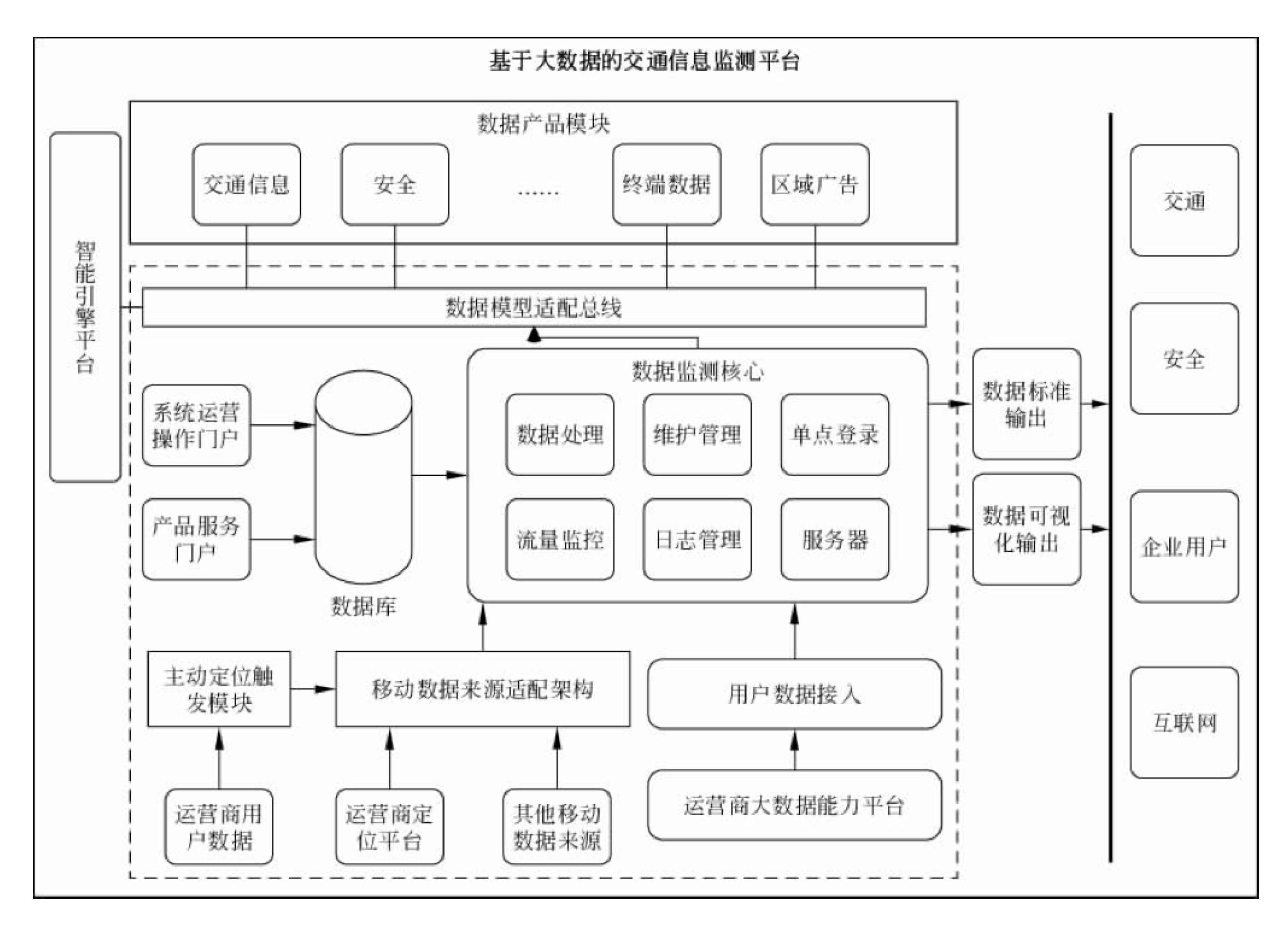


图 8-8 基于大数据的交通信息监测平台

预测，并在不断积累数据的过程中不断优化预测分析结果，提升平台的计算分析能力。

基于此的交通数据信息数据处理流程如图 8-9 所示。

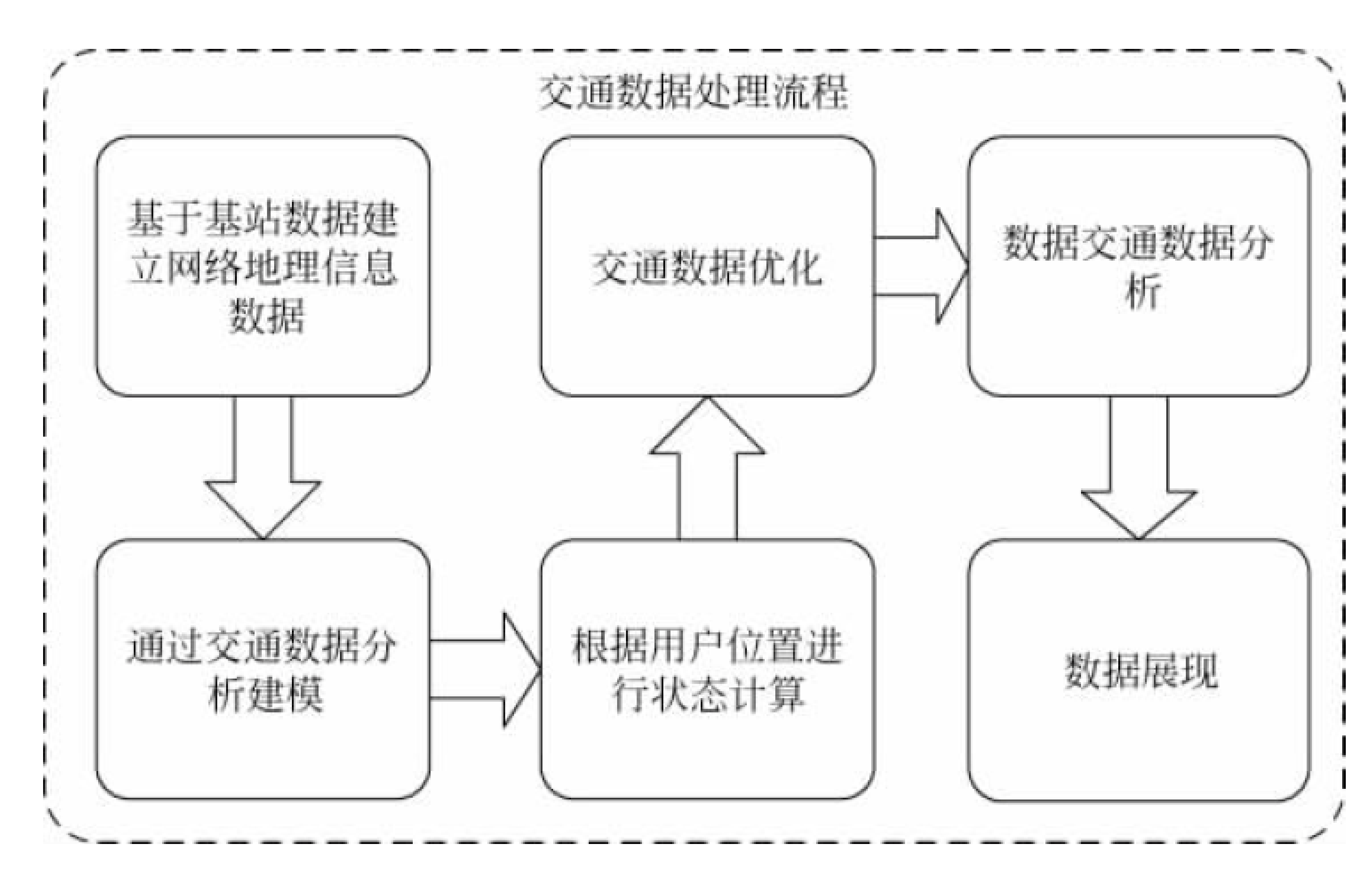


图 8-9 交通数据信息数据处理流程

交通信息数据获取平台可以获取到当前交通网络上有效用户的实时交通数据，利用用户速度算法模型，得出每个用户的实时速度数据，并可以对当前区域的用户的速度标签做多源优化，输出区域交通速度数据，进而通过可视化模块对城市的不同区域的道路信息实现可视化实时监控，为用户提供实时的数据服务，结合智能导航算法或相关工具，提供更优质的

出行方案。同时还可以结合现有数据实现交通运行情况的历史数据对比分析和实时查询，为政府智慧城市、交通等相关部门提供专业的、准确的数据服务。

通过建立交通信息监测平台，根据不同的使用需求，可以在现有的摄像头为主要途径的交通监测手段的基础上，从另一种途径实现实时的交通信息发布和道路交通情况的监测；又可以通过对历史数据的积累，研究其变化趋势，对未来的发展情况形成预测；也可以用来分析重点城市道路密度、汽车保有量、人口、节假日交通状况、城市主要交通状态等方面；还可以为未来政府智慧城市的建设提供专业、准确的数据、信息服务。

该通信企业基于大数据的交通信息监测平台也具有如下优势和特点。

（1）海量的、多维度的交通数据；

（2）高精度的地图数据；

（3）坚实的硬件和数据网络支持；

（4）高性能的大数据的分析技术；

（5）多种数据可视化的呈现；

（6）平台提供易用的开发应用编程接口及开发套件，支持多种交通信息技术指标。

通信企业依托自有的硬件、基站、用户数据和技术优势，建立起基于大数据的交通信息监测平台，从众多的信息源获取实时数据，通过对海量数据的挖掘分析，实现实时、高效的交通信息监测，通过不同模块实现对数据价值的有效发掘，进而对交通、安全等多个领域提供助力。

在这个案例中，需要实时从网上获得海量的移动用户的信息、地图信息、交通台信息进行计算，并要存储统计得到的历史信息。数据分布广泛、数据体量大、数据实时性与计算速度的要求很高，这些要求只有在大数据技术的帮助下才能完成。

第四篇　大数据，推动新型政务

随着新一轮的信息技术与产业、经济和社会的深度融合，大数据日渐成为社会发展的战略性资源。工业和信息化部赛迪智库网络空间研究所所长刘权认为，大数据是“未来的新石油”，一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，未来对数据的占有和控制甚至将成为继陆权、海权、空权之外国家的另一个核心资产。

面对海量、动态、多样的大数据，传统的思维方式和行为方式面临巨大挑战，尤其是在公共服务领域。大数据时代，如何推进社会治理与政府职能转型，提升政府治理和公共服务能力和水平，成为一个全新的课题。

中国计算机学会(CCF)大数据专家委员会秘书长程学旗表示，大数据运用有助于推动政府治理理念和模式，进而加快治理体系和治理能力现代化，同时也为推动政府治理决策精细化和科学化带来机遇。在大数据时代，海量数据能够对经济社会运行规律进行直观呈现，使政府治理所依据的数据资料更加全面，从而降低政府治理偏差概率，提高政府治理的精细化和科学化。



大数据时代的税务精细化管理

大数据技术在经历概念、炒作、沉淀之后，当下正在逐步应用到各个行业中解决实际问题，例如，电信行业的详细通话记录查询、电商行业的商品推荐、金融行业的风险防控等，当然还有很多其他的例子，可以说大数据技术已经在我们看不见的地方影响着生活的方方面面。数据也有属于自己的基因，它在不同行业中展现出来的特性是不同的，对其价值的挖掘必须要遵从其本身的特点和规律。

税务机关作为国家重要的职能部门，是国家财政收入支出体系中最重要的一环，由于其工作内容、性质和重要性，在早年就被国家纳入了“十二金”工程，其信息化程度在政府职能部门中一直排在前列。随着信息化的不断深入，税务机关的数据在数量、种类、价值、处理速度上已经满足了大数据的标准，换句话说，数据环境已经逐渐成熟，具备应用大数据技术的条件。

9.1 大数据时代下税务工作新趋势

9.1.1 税务数据新趋势

税务工作的主要任务是为民执法，为国聚财，参与宏观调控国民经济。这一切的实施依赖于对纳税人按照不同税率进行准确的征收工作。如何合理地制定不同的税率和进行高效的征收，在已有税务工作经验的基础上，更需要对税务数据进行分析，找出背后的规律，用规律科学地进行指导。

可以说未来的税务工作应该更多地围绕税务数据展开，然而在云计算、大数据、互联网＋推动的信息化浪潮下，税务数据本身也发生了变化，变化趋势如图 9-1 所示。



图 9-1　税务数据的变化趋势

1. 税务数据量的增加

税务工作本身就是一个复杂的系统，涉及实体众多，包含税务机关、企业、事业单位、个体工商户、单个自然人等。按照税种又分为营业税、房地产税、个人所得税、车船税等。实体和税种之间是多对多的关系，管理这些多对多的关系本身就将产生大量的数据。

随着国家大力推进实体经济的发展，越来越多的单个自然人参与其中，在各行各业从事着种类繁多的经济活动，新的经济形态层出不穷，Uber、Airbnb 等代表的个人经济活动成为越来越普遍的情况，导致涉税实体之间的关系越来越复杂，在如此复杂的关系网中，针对如此频繁的个人经济行为进行记录和追踪，由此带来爆炸式的数据增长，如何从税务工作角度进行管理是未来税务部门必须要解决的问题。

2. 税务数据种类的增加

在未来，特别是互联网＋行动进一步推进的时候，运用传统的税务管理思维来开展税务工作已经无法适应新时代的需求。在数据层面上，税务数据应该走出去，和外部的单位和企业开展合作，积极地引入第三方数据，用来很好地补充税务数据的单一性。

比如在核算个体工商户应纳税额的环节，如果能有该纳税户的月交易数据（刷卡交易）和互联网对其的评价和关注度，就能很准确。同时税务数据在其他领域也可以发挥很大的作用，比如将企业的纳税情况作为参考提供给征信领域使用，当前已经出现的税银贷就是相应的跨部门合作的新形势。

外部数据的引入除了在量上带来增加外，在数据种类上也会更加丰富。从单一关系型数据库的结构化数据到类似 XML 的半结构化数据，再到声音、图片等非结构化数据，今后都会是税务工作所需要的。

3. 税务数据处理速度的加快

税务数据不仅在量上和结构上发生变化，对其处理速度也会有更高的要求。原因有三点，一是数据量和种类的增加本身就会带来处理速度的下降，所以要提速；二是在应对突发事件和监控类应用时，时效性是重要的性能指标；三是数据分析类应用只有在快人一步的情况下才能发挥最好的作用，先做出正确决策的人往往都是赢家。

9.1.2 税务业务新趋势

1. 智能业务

税务工作在经历多年信息化建设后，已经大大提高了办税效率和税收服务效果，但是在经济活动种类和数量不断增加的今天，还是面临着偷税漏税、纳税渠道不畅、税收风险控制等诸多应用上面的难题，而大数据将成为解决这些问题的新思路和新办法。大数据的解决指导是引领智能税务的出现，其本质是用大数据技术找出数据背后的规律和关联关系来解决税务工作中的难题。

面对这些应用难题，大数据从底层税务数据开始着手，经过对海量税务数据的清洗、加工、转换形成有价值的税务数据资产，然后利用算法模型来发现数据的规律和关联关系，并结合税务工作中难题的业务需求，最终形成数据应用来解决它。由于在过程中运用了大量基于事实的数据和先进的算法模型，这些应用将会成为税务工作迈向智能化的关键助推器。

2. 业务共享

大数据在税务行业的应用，除了推动税务智能化进程外，对工商、公安等其他政府职能部门和广大纳税人也是有很大好处的。

除了对税务部门的帮助，在税务工作中引入大数据，还能通过信息互联、数据共享惠及如工商、公安等其他政府职能部门。当然纳税人也能因此获得更好的纳税服务体验。大数据技术通过税务数据应用将税务部门、其他职能部门、纳税人连接起来，形成联动效应，以发挥税务数据的最大能量。

一方面通过大数据税务数据应用产生的税源评估可以为工商注册审核、公安经济犯罪案件侦破起到辅助作用，帮助锁定有问题企业和个人；另一方面通过优化纳税渠道和纳税流程，让纳税人享受到更便捷的纳税服务。

9.1.3 新机遇和新挑战

税务系统的数据在很长时间内大量来自于纳税人的申报行为数据和报表数据，面向税务工作人员的是割裂的不同业务系统，信息本身被业务消解为固定的逻辑和处理形式，这样的工作形式能够直接服务于不经常变化的业务形态，但看似简单的数据口径，一方面隐藏了大量的数据细节，另一方面也给快速变化适应带来了现实的困难。

当前的税务部门需要适应不断涌现的各种经济形态，从活跃已久的电子商务到方兴未艾的个体经济（Airbnb，Uber），都需要税务工作者快速响应，根据数据进行记录观察和监督管控。业务逻辑层在不断简化变薄，工作人员会发现大量需求落实到对数据本身的探求和感知，灵活的数据查询需求早已不能够被僵化割裂的业务系统满足。

在架构方面，数据的积累在漫长的信息化过程中迎来了爆发增长的时期，各种综合治税和第三方数据进入税务工作的视野。传统基于单机的处理架构开始日益缓慢和臃肿，几乎所有建立在数仓之上的应用，都开始面临扩容升级的不菲成本。但令人头痛的事并不在于增长的数据，而在于数据的增长速度，升级后容纳一个新的数量级不是问题，但想要频繁地持续地扩容升级，则意味着必须要在基础架构上做出选择。

在应用方面，税务行业在历经单机查询，省级集中征管系统，数据仓库等信息化发展历程后，如今迎来更进一步的数据融合、数据智能阶段。数据成为最重要的信息资产，需要有效发现、收集、管理、分析。从数据的视角看，税务部门的考察对象是税源，即合法纳税人的信息真实性。在传统税源管理工作中，受限于纳税人申报信息的有限性和滞后性，税务工作者无法准确把握纳税人的全面情况。得益于日趋成熟的互联网技术，一方面互联网的很多公开权威信息能够形成信息的互通互联相互比对，另一方面基于互联网迅猛发展应运而生的大数据技术服务，能够提供包括风险建模、预测分析、关联分析等高级数据应用管理工具。使用大数据技术进行税源专业化管理，有利于建立以税源为中心、比传统业务模型更加强大和全面的综合信息视图，并以此为依托进一步搭建包括税源关联分析、分类画像、风险预测在内的一系列数据应用。

总的来说，税务机关在架构和应用两个方面都遭遇了挑战，究其原因还是由于新形势下对税务工作提出的新的要求。在大数据时代，税务工作要完成从面到点、从粗放到精细、从人管税到数管税的转变，对税源实行精细化管理。

2. 业务共享

大数据在税务行业的应用，除了推动税务智能化进程外，对工商、公安等其他政府职能部门和广大纳税人也是有很大好处的。

除了对税务部门的帮助，在税务工作中引入大数据，还能通过信息互联、数据共享惠及如工商、公安等其他政府职能部门。当然纳税人也能因此获得更好的纳税服务体验。大数据技术通过税务数据应用将税务部门、其他职能部门、纳税人连接起来，形成联动效应，以发挥税务数据的最大能量。

一方面通过大数据税务数据应用产生的税源评估可以为工商注册审核、公安经济犯罪案件侦破起到辅助作用，帮助锁定有问题企业和个人；另一方面通过优化纳税渠道和纳税流程，让纳税人享受到更便捷的纳税服务。

9.1.3　新机遇和新挑战

税务系统的数据在很长时间内大量来自于纳税人的申报行为数据和报表数据，面向税务工作人员的是割裂的不同业务系统，信息本身被业务消解为固定的逻辑和处理形式，这样的工作形式能够直接服务于不经常变化的业务形态，但看似简单的数据口径，一方面隐藏了大量的数据细节，另一方面也给快速变化适应带来了现实的困难。

当前的税务部门需要适应不断涌现的各种经济形态，从活跃已久的电子商务到方兴未艾的个体经济（Airbnb，Uber），都需要税务工作者快速响应，根据数据进行记录观察和监督管控。业务逻辑层在不断简化变薄，工作人员会发现大量需求落实到对数据本身的探求和感知，灵活的数据查询需求早已不能够被僵化割裂的业务系统满足。

在架构方面，数据的积累在漫长的信息化过程中迎来了爆发增长的时期，各种综合治税和第三方数据进入税务工作的视野。传统基于单机的处理架构开始日益缓慢和臃肿，几乎所有建立在数仓之上的应用，都开始面临扩容升级的不菲成本。但令人头痛的事并不在于增长的数据，而在于数据的增长速度，升级后容纳一个新的数量级不是问题，但想要频繁地持续地扩容升级，则意味着必须要在基础架构上做出选择。

在应用方面，税务行业在历经单机查询，省级集中征管系统，数据仓库等信息化发展历程后，如今迎来更进一步的数据融合、数据智能阶段。数据成为最重要的信息资产，需要有效发现、收集、管理、分析。从数据的视角看，税务部门的考察对象是税源，即合法纳税人的信息真实性。在传统税源管理工作中，受限于纳税人申报信息的有限性和滞后性，税务工作者无法准确把握纳税人的全面情况。得益于日趋成熟的互联网技术，一方面互联网的很多公开权威信息能够形成信息的互通互联相互比对，另一方面基于互联网迅猛发展应运而生的大数据技术服务，能够提供包括风险建模、预测分析、关联分析等高级数据应用管理工具。使用大数据技术进行税源专业化管理，有利于建立以税源为中心、比传统业务模型更加强大和全面的综合信息视图，并以此为依托进一步搭建包括税源关联分析、分类画像、风险预测在内的一系列数据应用。

总的来说，税务机关在架构和应用两个方面都遭遇了挑战，究其原因还是由于新形势下对税务工作提出的新的要求。在大数据时代，税务工作要完成从面到点、从粗放到精细、从人管税到数管税的转变，对税源实行精细化管理。

9.2 大数据技术的价值

1. 大数据技术的核心

大数据技术的核心是大数据存储和处理技术、数据仓库技术等，其战略意义在于掌握和处理庞大的数据信息。大数据应用的核心是实时数据处理、实时决策支持，其战略意义在于快速地分析出数据的价值，让价值发生作用，通过内嵌到业务流程中实现数据价值的体现。也就是说，大数据应用的核心价值。

2. 提升数据分析整体效能

大数据应用的实时处理、实时支持、内嵌流程的要求，是区别传统数据分析利用的关键差别之一。大数据应用的目标，是支撑所有税务工作人员的实时业务处理和日常管理需求，而并非仅仅是管理层的分析统计和决策支持需求。

3. 改变传统的数据分析

大数据应用提供嵌入业务流程的决策支持能力，提升日常管理的决策效果。如果不能从流程的观点考虑问题，大数据应用可能只能提供一些相互割裂、为局部目标服务的独立的数据分析应用，只能达到局部优化的目的。大数据应用可以是流程中的黑盒子，整合在业务流程之中，无论有没有高深的数学算法和统计模型，整个业务流程都能够运转，而当有更好的、通过验证的算法出现并融入到流程时，整个流程的绩效将得到提升。所以，大数据应用不应该仅关注数据分析的角度，而是支撑整个税收征收管理的优化和演进。

4. 研发大数据应用

传统的数据分析应用是按照提出假设、发现模型、选择数据、建立关联、定义模型的模式进行设计。而大数据应用采用的不是随机分析法（即抽样调查）这样的捷径，而采用对所有数据进行分析的方法。特别是基于电子（网络）发票明细数据的分析，以及来自互联网的业务数据的分析，对于整体计算存储能力提出了更高的要求。通过应用云计算技术，可以整合税务系统内部在地理上分散的计算存储资源，还可以实现按需从公共云计算平台中采购相应的计算能力，或根据业务需要扩展订购、更换更加适合的应用系统服务。

5. 大数据应用的监控

大数据应用不是毕其功于一役的运动式项目，而是应该能够对税务管理做出持续的改善。如果没有持续的监控，大数据应用所带来的改善可能会很快消失，税务机关将退回到项目开展前的状态。因此，大数据应用要详细设计监控的方法和流程、数据获取的方式、度量绩效的监控指标和方式，并且能够以绩效仪表盘的方式将监控结果以可视化的方式展现给用户，以利于大数据应用效果的可持续性。

6. 区别于传统数据仓库

传统的数据仓库是数据驱动，主要关注对于已经掌握的海量数据的建模、处理和分析。大数据应用以业务问题为主要驱动，从管理需求出发，通过主动寻找新的数据来源、设计更好的人机交互方式、设计实验和验证等方式更加主动地搜集数据，以获取为支撑决策所需要的数据和证据。因此，大数据应用的要点在于：一是如何整合不同来源的数据（如企业财务报表、税源管理数据等）并建立关联关系；二是如何帮助业务人员方便灵活地获取所需粒度

的数据进行即席分析，以应对管理环境变化而做出权变的决策；三是更加关注于数据获取方式、模型动态选择、业务规则和业务逻辑的管理，重点分析并掌握纳税人的“行为指纹”，以“沙盘推演”的方式帮助税务机关事前选择和事中优化管理决策行为；四是在为用户提供便捷数据访问的同时，更关注于分析结果的权限受控。因此，基于大数据应用提升经验总结并形成对于税务管理的洞察力，并运用计算机系统进行自动推理，是大数据应用最有价值的内容。

9.3　税务精细化管理顶层设计

税务的精细化管理依赖于对税务数据价值的利用，如何系统地、高效地利用数据需要有方法论的指导，需要做好顶层设计，如图 9-2 所示。

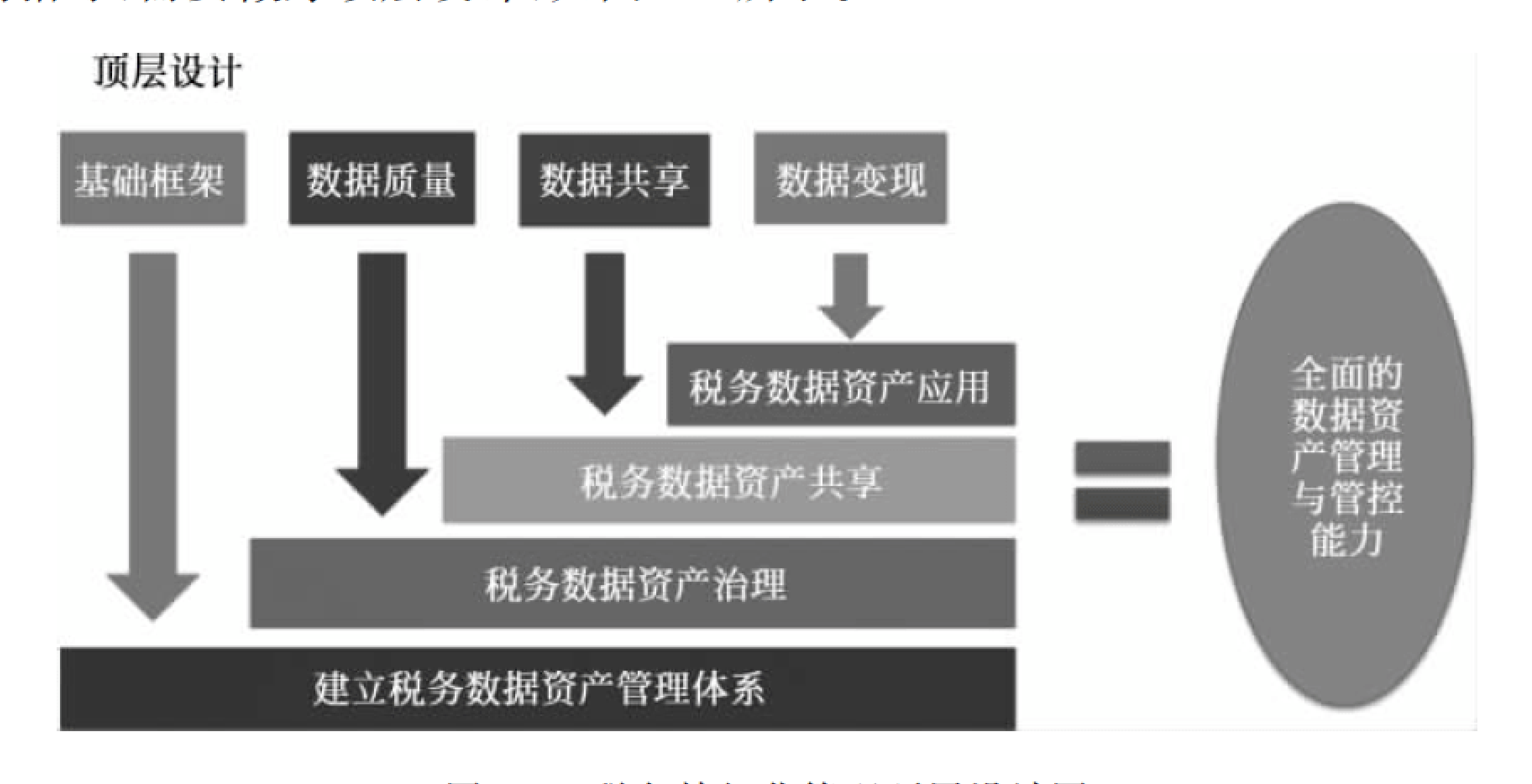


图 9-2　税务精细化管理顶层设计图

通过制定由数据资产治理、数据资产运营、数据资产应用三部分来组成总体的数据资产管理体系框架。

通过制定与数据资产相关的一切标准和处理流程，一方面把控数据质量，一方面规范处理流程来形成数据资产治理体系。

通过数据接入和分发来打通税务数据和外部数据，形成多层次、多维度的税务大数据共享环境。

通过灵活、敏捷地运用并行化模型算法开发具有实际业务含义的大数据应用，真正让数据体现出价值。

有了明晰的数据资产管理体系顶层设计，还要配套制定出切实可行的实施步骤，按照步骤分阶段建设，最终达到基于税务数据资产利用的精细化管理。整个税务资产管理总共分为 5 个阶段，如图 9-3 所示。



图 9-3　税务资产管理总 5 阶段

具体包括基础能力建设、数据汇集、数据治理、数据应用、数据运营。其中，基础能力建设是指通过构建 Hadoop 集群来提供数据存储和数据处理的能力；数据汇集是指将散落在各个系统中的结构化数据、半结构化、非结构化数据都导入到大数据基础平台中；数据治理

是对汇集后的数据进行清洗和校验，提高数据质量，让数据变得真正可用；数据应用是指基于数据进行各式各样的应用开发，从基本的数据查询到数据的多维度统计分析，再到利用算法模型进行数据挖掘等；数据运营是指利用数据资产进行数据开发、数据交易和数据合作。

9.4 大数据税务应用整体架构

9.4.1 总体架构

在整个税务精细化管理过程中，税务数据应用是手段和工具，同时也是数据最终变现和发挥价值的地方，是关键所在。围绕税务数据应用将整体逻辑架构设计为如图 9-4 所示。

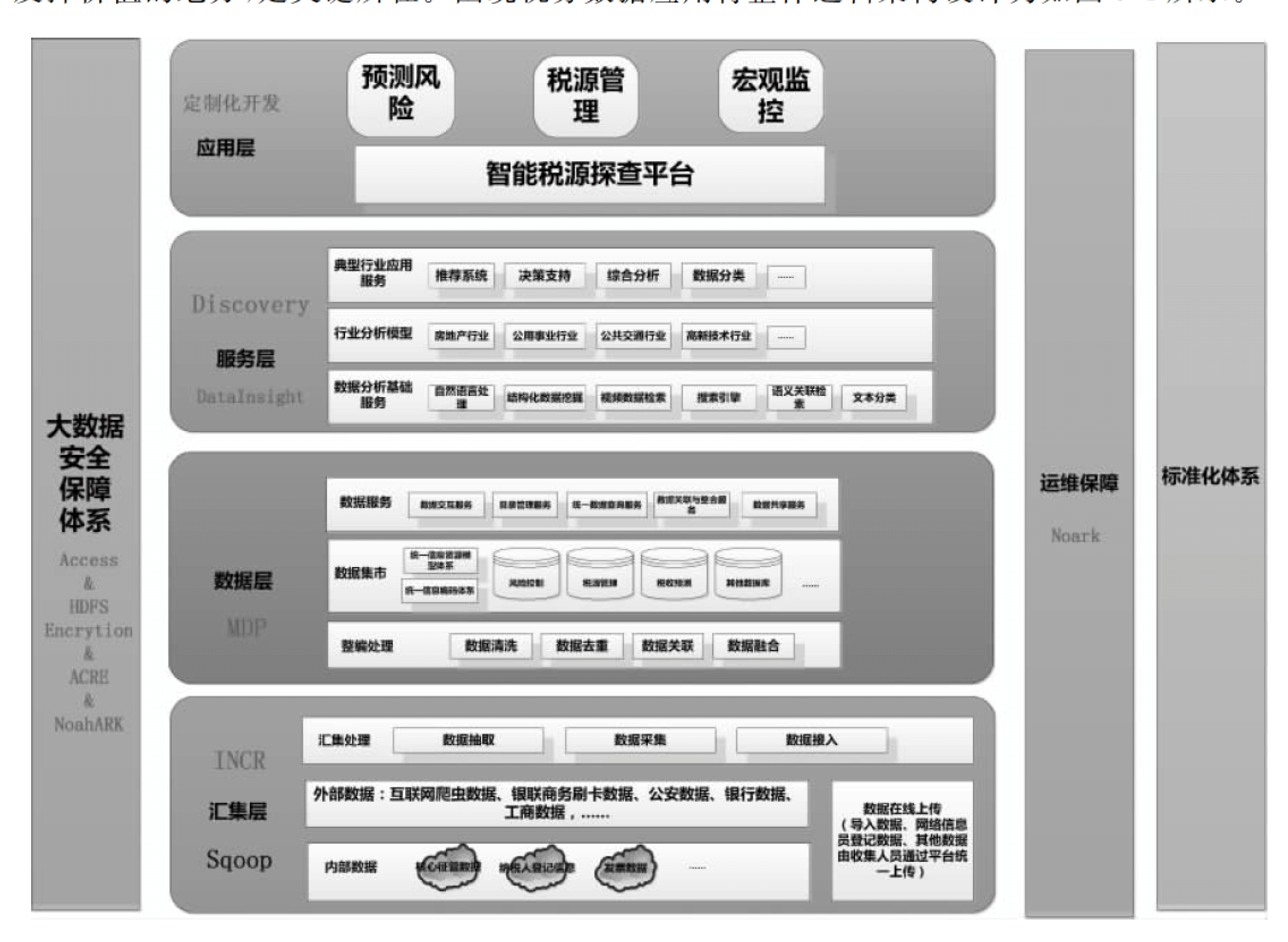


图 9-4 税务数据应用整体逻辑架构设计

从图 9-4 可以看出整个逻辑架构从最底层的汇集层开始，往上分别是数据层、服务层以及应用层，辅以大数据安全保障体系、运维保障和标准化体系。目的是涵盖税务数据全生命周期的各个环节，做到逻辑清晰、层次分明，同时满足系统对安全、运维、标准化方面的要求。

总的来说，原始数据经由汇集层聚集到平台中，然后通过数据层转换为高质量的数据分析挖掘源，再经过服务层提供的模型算法挖掘能力，最终在应用层通过数据应用的方式实现数据价值的变现。

9.4.2 汇集层

汇集层是为解决数据来源的问题。按照内部数据和外部数据的界面，将分散在不同职能业务部门的业务子系统中的业务统计数据抽取出来，以规范的数据模型组织并存储在数

据平台中，形成统计信息数据源。根据数据存在的形式又可以将系统的数据分为以下三类。

（1）结构化：传统的关系数据模型、行数据，存储于数据库，可用二维表结构表示，例如，核心征管业务信息数据库、纳税登记信息数据库。

（2）半结构化：类似 XML、HTML 之类，自描述，数据结构和内容混杂在一起，例如，电子政务办的网页信息、网站的配置文件等。

（3）非结构化：各种文档、图片、视频/音频等，例如，Word 文档存储的视频数据、高清地图数据等。

针对将标识出的数据源以及数据源的分类，采取不同的获取策略，设计通用的模块，进行数据的获取。将数据获取层获取的数据，根据具体的应用场景，采用不同的导入策略，将数据导入到数据层。

9.4.3　数据层

数据经由汇集层存储到平台中后，在汇集层将数据汇集到数据层的基础之上进行各种加工，例如，数据去重、数据一致性检查、数据标准化和数据的格式化转换。

（1）数据清洗：数据清洗的目的就是按照一定的规则统一数据格式，过滤脏数据，保证后续过程的安全稳定运行。

（2）数据去重：数据去重的目的就是将原始数据中重复出现的字段或数据去掉，降低数据的冗余度。

（3）数据融合：数据融合的目的就是按照一定的业务规则，将不同维度的原始数据组合起来，形成新的有实际业务含义的数据集。

（4）数据关联：数据关联的目的就是按照某些字段，将分散在不同原始数据源的数据关联起来，得到更完整维度的数据。

这些操作是为了保证基础数据的质量，为上层的应用提供可靠的数据支撑。这一层中存放的数据是具有核心价值的数据，它将为数据分析系统提供最基础的数据，以深入地挖掘数据的价值。数据层主要存储数据和任务相关的元信息，以及数据元层中的各种数据。

根据不同的业务需求在核心数据之上可以构建专题数据仓库，例如风险控制数据仓库、税收预测数据仓库、税源管理数据仓库等，来满足专题研究分析数据源的需求。同时从数据本身出发，对上层提供数据交互服务、目录管理服务、统一数据查询服务、数据关联与整合服务、数据共享服务等一系列服务。

9.4.4　服务层

服务层旨在提供各类基于数据的基础服务，为上层应用开发提供便利。

首先，服务层提供包括自然语言处理、结构化数据挖掘、视频数据检索、搜索引擎、语义关联检索、文本分类等一系列数据分析挖掘技术和算法，用户直接调用即可完成相应的数据分析，而不必自己面对复杂的算法并行化实现。

其次，根据业务需求，服务层建立了以房地产行业、公用事业行业、公共交通行业、高新技术行业为代表的一系列行业分析模型，目的是将常用模型进行总结和归纳，得出较为通用的部分模型，缩减用户建模挖掘的时间，为用户提供便利。

最后，基于数据分析基础服务，服务层还提供了诸如推荐系统、决策支持、综合分析、数

据分类等一系列典型行业应用服务。

9.4.5 应用层

应用层是展现给最终用户使用的，是整个项目的数据变现出口。首先构建了智能税源探查平台，用户通过登录这个平台就能满足其所有对数据的需求，包含数据的存储、查询、分析、挖掘、展现等。

智能税源探查平台提供预测税收风险、税源精细化管理、税收宏观监控等数据应用。目的是通过这些应用持续完善税收征管体系、优化纳税渠道、建立面向涉税数据全生命周期的监控体系，探索智能化的纳税服务。

9.4.6 大数据安全保障体系

税务数据是高安全级别的敏感数据，它的安全至关重要。智能税源探查平台会以 Hadoop 生态系统为核心进行建设，所以 Hadoop 生态系统的安全就决定了整个平台的安全。我们需要从安全需求出发对 Hadoop 生态系统进行全面的安全加固。

Hadoop 生态系统是数据储存和数据处理的实体，由很多的组件构成，不同的组件负责具体的功能，比如 HDFS 负责数据储存、MapReduce 负责数据处理。组件和组件之间，组件和外部系统均存在联系，如图 9-5 所示。

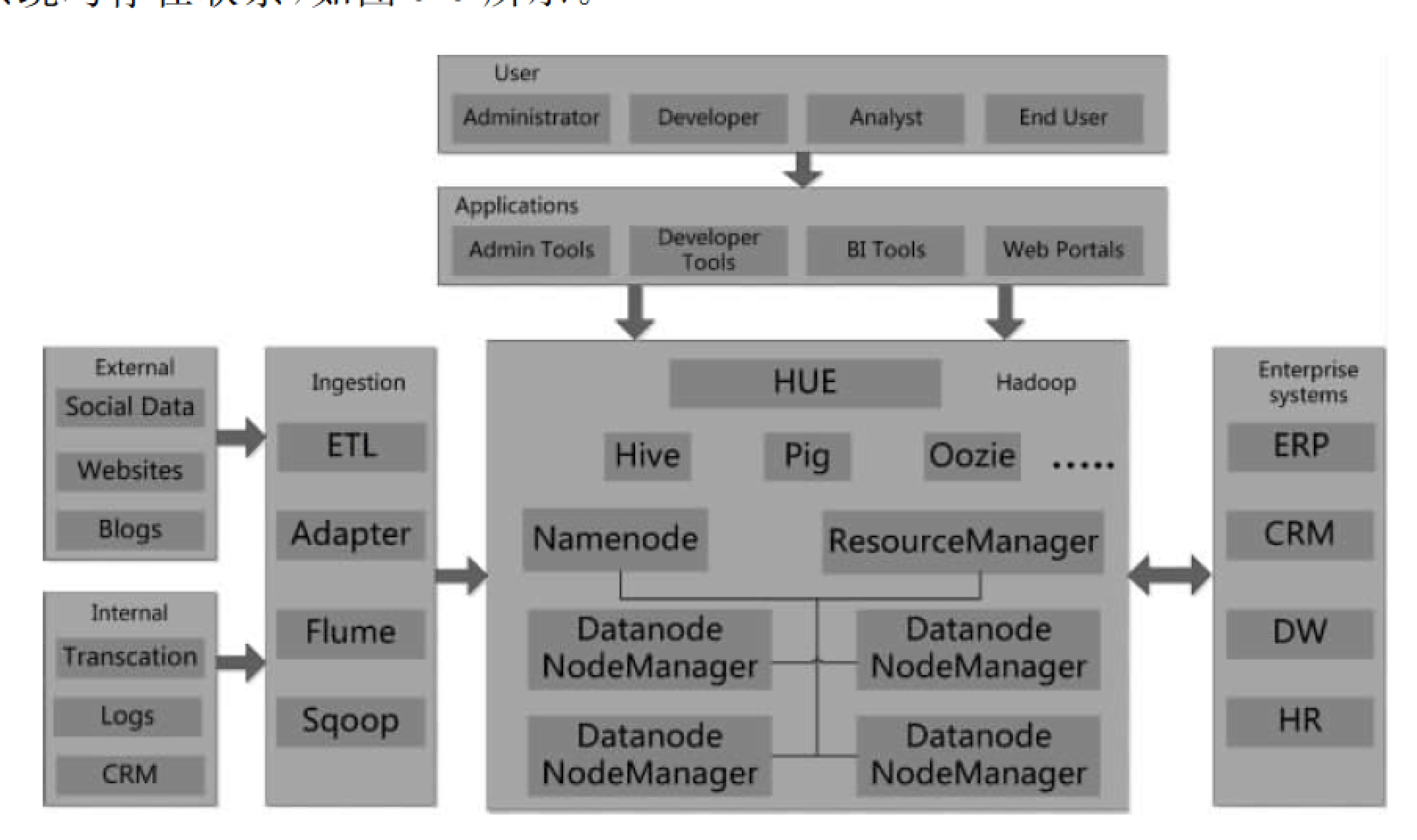


图 9-5 Hadoop 生态系统

可以看出整个 Hadoop 生态系统的安全是比较复杂的，要从组件服务和数据生命周期两条主线下手，具体来说应该从以下几个方面去考虑。

(1) 认证：需要为用户和服务提供统一的认证机制，确保没有未经认证的第三方接入平台获取数据和服务，并且该认证机制要能和现有的用户管理体系相结合。

(2) 授权：需要提供基于角色的授权机制，确保用户只能够进行与其角色身份相符的操作和相应的集群资源等。

(3) 审计：需要提供针对数据变化和用户操作的记录手段，用以在发生问题的时候追溯责任、分析原因以及预警。

(4) 数据加密：需要提供合适的数据加密方法，对存储的静态数据和传输的动态数据

进行加密保护。

(5) 数据传输：需要提供数据在 Hadoop 平台和外部系统间流动时的安全机制，保证数据在传输过程中不发生泄露，保证 Hadoop 平台一定的独立性，防止脏数据流入。

9.4.7 运维保障

智能税源探查平台的运维主要分为软件平台运维、数据资产运维、数据安全运维。

其中，软件平台的运维着重在对各系统规范化要求、常用软件的安装和部署、版本控制、权限控制、备份控制、数据控制、实时监控报警等方面发挥作用。

数据资产运维着重在数据生命周期管理、任务调度管理、数据质量管理等方面发挥作用。

数据安全运维基于平台产生的各类日志文件，通过日志收集、异常规则、异常发现、规则训练、异常预警等步骤在数据安全方面发挥作用。

9.4.8 标准化体系

标准化体系建设是为将来与各个共建单位进行对接而准备的，需要在平台建设的同时就考虑到各类数据和应用的调用接口，实现接口的标准化和通用化，在平台对接过程中做到平滑、无缝。

具体做法就是在接口开发过程中，尽量采用业界通用的协议或者标准，如 RESTFUL API、JDBC 连接、FTP 文件传输协议等。如果有特殊情况，需要开发个性化的接口，则要提供完备的对接支持，包含文档和技术支持两方面。

9.5 大数据税务数据应用场景

大数据技术在税务领域发挥作用，需要和税务工作人员进行密切沟通，从中梳理出税务工作中的痛点，分析其中的关键问题。然后根据已知内外部数据的情况，寻找合适的算法模型来建模，最终提出解决方案。

税务数据应用总体架构中最上层列举了预测风险、税源管理、宏观监控三个大方面的应用场景，我们将税务数据应用场景进行了细分，归纳起来有以下 5 个部分的内容。

一是需要针对纳税渠道进行资源的差异化配置和补给，用以提升优质纳税渠道的顺畅度，使得纳税人更加便捷，提高办税效率；二是需要发现企业间潜在的社会经营关系，用以检测两者间的生产经营关系是否存在问题；三是需要结合外部数据对深层次的偷税漏税问题进行跟踪和追缴；四是需要思考大数据在税务领域的创新型应用；五是需要了解纳税人全方位的信息，用以制定差异化的管理措施，对税源实现精细化管理。

税务数据应用的重点落在第 4 部分即大数据在税务领域的创新型应用即涉税事件追踪分析上。下面就每个具体的应用场景从场景描述和解决方案两个方面来进行详细叙述。

9.5.1 优化纳税服务

1. 场景描述

针对不同的纳税渠道每年都会有资源投入来进行维护和拓展，但是由于无法区别出纳

税渠道的优劣，在资源分配时只能采用均摊的方式，导致优质渠道得不到足够的资源，服务体验不好的渠道浪费资源，从而使得纳税人的纳税体验无法得到提高。

针对上述场景，我们认为可以通过对数据的分析得出各个办税渠道的用户量和顺畅度情况，以此来指导国税局对办税资源的差异化配置，提高办税资源的利用率。

2. 解决方案

各个纳税渠道的优劣可以通过打分的方式来评定，我们使用加权平均算法来完成这项工作。该算法的思想是挑选出评分目标的关键属性值，然后为不同的属性值赋予不同的权值，最后将所有属性值乘以权值并进行求和，运算结果就是评分目标的最终得分。

我们首先梳理出与纳税渠道属性相关的数据源，如国税局目前所有的办税渠道、纳税人网页上的行为数据、纳税人的网络评价、纳税人常用的纳税渠道、纳税的频率、是否有欠税情况、缴纳时间离税款所属期开始时间最近的次数、提前缴纳税款的次数等。

然后从上述数据中经过转换和计算得出纳税渠道关键的5个属性值，一是统计国税局某一时间段内所有办税渠道的纳税次数；二是搜集纳税人对各个办税渠道的评价，转换为从1到5的评分；三是统计各个办税渠道缴纳时间离税款所属期开始时间小于5天的纳税次数；四是统计各个办税渠道税款欠缴的次数；五是统计各个渠道纳税人从申报开始到最终拿到完税证明的时间。

设 X_1、X_2、X_3、X_4、X_5 代表5个关键属性值，相应的权值为 W_1、W_2、W_3、W_4、W_5，那么根据如下公式：

$$\bar{x}=\frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n}$$

计算出各个纳税渠道的最终得分，因为在数学中加权平均数比起简单的算术平均数更能反映评分目标在某一方面的变现，所以依据最终结果由高到低来为各个办税渠道差异化地配置资源是科学可行的。

3. 思维导图

将上述解决方案以图形化的方式展现出来便得到如图9-6所示的图形。

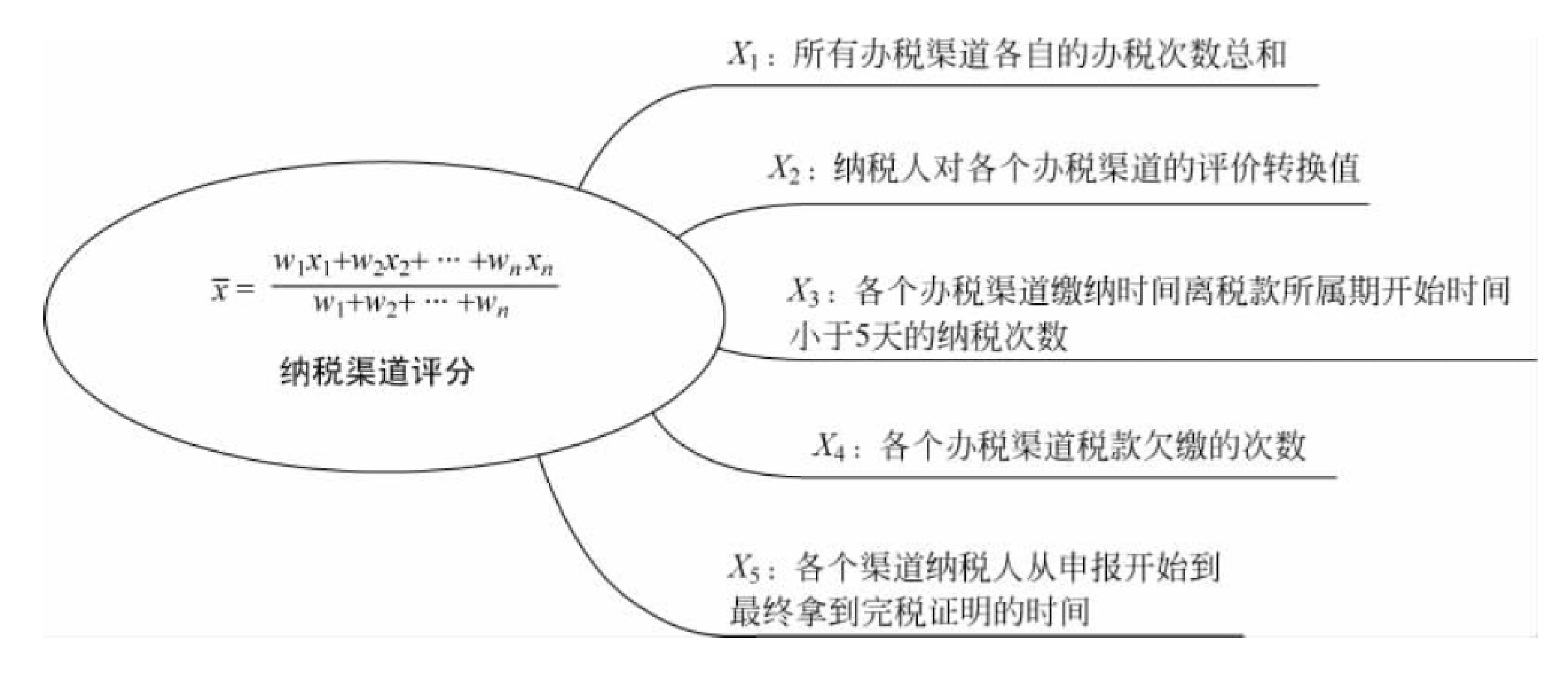


图9-6 优化纳税服务思维导图

9.5.2 社会经营关系

1. 场景描述

在进行税源管理时，仅仅能从交易往来发票中得出一些企业间的生产经营关系，但是无法判断这些关系的真实性和信息量。企业间会通过各种手段来隐瞒彼此之间的真实经营往来情况，导致虚开、走票、少开等违规违法现象的出现，造成税收收入的损失。

针对上述场景，我们认为通过搜集和分析企业间的社会关系能判断出企业间的生产经营关系是否如其在国税业务系统中的数据所反映的那样，从而避免由于虚假生产经营关系造成的损失。除此之外，企业间的社会关系还能对纳税稽查、税源审核等工作有所帮助，在税源管理中扮演着十分重要的角色。

2. 解决方案

企业间的社会关系可以通过分析企业管理人员和高级员工间的关联关系得出，因为企业和企业间的生产经营往来，必然会带来人员的交流沟通，尤其在管理人员和高级员工之间，而且这两者之间是明显的线性正相关。通过比较企业的社会关系和生产经营关系，很快就能发现问题。

我们首先要搜集目标企业的注册信息、企业的法人信息、合伙人信息、股权结构信息等，从中识别出企业的管理人员和高层员工。但是仅识别出来依然无法判断企业间的社会关系。这里还需要引入外部数据，具体来说就是互联网数据和公安数据，互联网数据由互联网爬虫程序搜集，公安数据由相关合作方提供。

互联网数据是以企业管理人员和高层员工的名字和公司名称为关键字爬取的，爬取回来的网页内容涉及流媒体数据、新闻数据、社交数据等，通过对网页内容的分析就能大致得出这些人之间是否来往密切。而公安数据起到的作用是最大的，我们以相关人员名字和身份证号来匹配他们是否籍贯一致、是否居住地邻近、是否频繁通话、是否乘坐过同一班次的火车或飞机、是否在同一时间入住过同一家酒店等。

为互联网数据和公安数据的每一项设置是或者否两个值，如果企业间两个相关人员在所有数据选项中超过60%都为是，那么就判定两人存在密切的社会关系。接着对筛选出来的所有人员进行一一匹配，形成一个 $N\times N$ 的矩阵。矩阵中的元素就是相关两人的社会关系，如果密切则标注为1，反之则为0。然后对矩阵中所有的元素进行求和，求和的结果如果大于矩阵元素个数的60%，则判定两家企业间存在密切的社会关系，反之则不存在。最后将结果与企业间的生产经营关系进行匹配，以发现问题。

3. 思维导图

将上述解决方案以图形化的方式展现出来便得到如图9-7所示的图形。

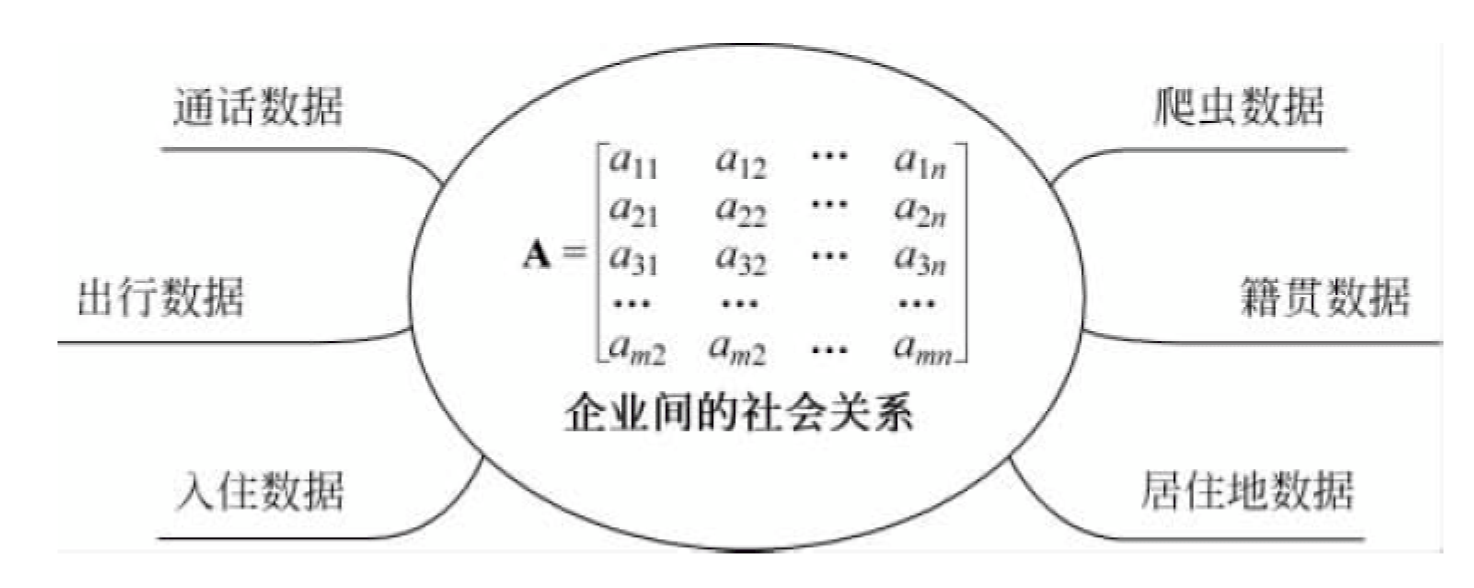


图9-7 社会关系思维导图

9.5.3 偷税漏税

1. 场景描述

偷税漏税一直是税务机关无法很好解决的问题，每年都会造成巨大的经济损失，损失一方面来自欠缴的税款，一方面是追缴税款所花费的人力成本和时间成本。企业偷税漏税现象的屡禁不止存在着主观意愿和客观条件两个因素。

其中，主观意愿是企业追求高利润的目的和侥幸的心理，在不少企业决策者眼里，纳税是一笔损失而不是每个公民的义务。他们更倾向于隐匿收入夸大成本，而达到不缴税和少缴税的目的。同时对于国家的法律惩罚，部分企业决策者对于偷漏税行为存在侥幸心理，在不完善的税收征管体系下，进行偷税的行为。

而客观条件是我国的税收管理体系存在漏洞，特别是在税款申报和税款核查环节。由于税务机关征收税款的直接依据是企业的账本，账本记录着企业的日常经济活动，但是现实生活中，企业能够通过多种手段来瞒报收入和经营状况，从而引发偷税和漏税问题。

针对上述场景，我们认为应该引入互联网数据、股票数据、银联刷卡数据等外部数据来对企业的生产经营和收入情况进行大致的预估，同时将企业历史同期值和行业平均值也纳入考虑范围，将上述估值和申报值进行比较，最终确定是否存在偷税漏税。

2. 解决方案

判断企业是否存在偷税漏税情况的关键所在是尽可能地摸清企业真实的生产经营状况和收入情况，然后将企业申报信息和分析结果进行对比。对比时要设定好一个阈值，如果两者的差异超过这个阈值就视为企业存在偷税漏税。所以阈值的设置比较关键，初期可以凭借经验，但是后期必须采用前期的反馈数据进行调整。

具体操作上，除了企业提供的财务信息，还需要搜集和分析企业的申报数据、企业的历史纳税信息、企业的股票数据、同行业的纳税数据、银联商务的刷卡数据以及互联网数据。首先采用逻辑回归算法对企业的历史纳税数据进行处理，通过在给定点集（各个时期的纳税金额）中拟合出一条曲线来预测企业未来的纳税金额。然后对该企业所在行业中所有与其规模相当企业的纳税额做一个算术平均得到该行业的平均纳税金额。接着如果在银联商务的刷卡数据中匹配到了该企业的数据，还需要根据刷卡产生的金额来折算出大致的营业额并按照税率换算成应纳税额。最后综合股票数据和以企业名称、产品为关键字爬取来的互联网数据对该企业目前的大致生产经营状况做出评估。

经过上述步骤得到了纳税额预测值 x_1、同行业纳税额平均值 x_2、纳税额折算值 x_3 以及生产经营状况评分值 x_4，这里还是用加权平均的算法对这 4 个变量求加权平均值，然后用这个值除以企业的申报纳税金额 x。计算过程如下所示：

$$\alpha = \frac{x_1 \times \omega_1 + x_2 \times \omega_2 + x_3 \times \omega_3 + x_4 \times \omega_4}{\omega_1 + \omega_2 + \omega_3 + \omega_4} \div x$$

如果 $\alpha < 80\%$，就代表企业申报额和估算值的相似度小于 80%，那么就可能存在偷税漏税的现象，需要进一步采取行政手段进行调查。80%就是前文提到的阈值，需要在实践中不断地进行修正。

3. 思维导图

将上述解决方案以图形化的方式展现出来便得到如图 9-8 所示的图形。

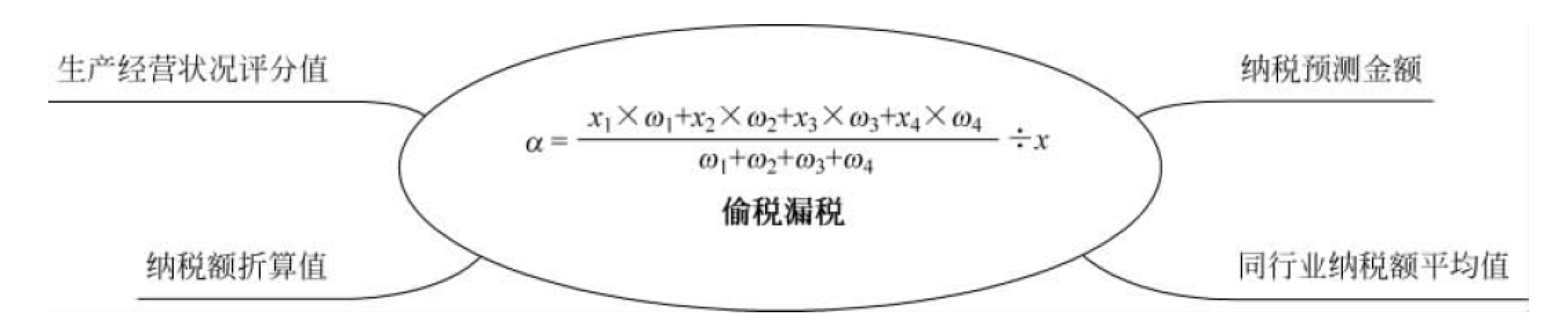


图 9-8　偷税漏税思维导图

9.5.4　涉税事件追踪

1. 场景描述

前文说到税务经过多年的信息化建设积累了大量的业务数据，但是这些业务数据是分散的、静止的，税务人员理解不了也利用不了。造成这种现象的根本原因是数据对于税务工作人员来讲是抽象的，只是一堆数字而已，无法与其日常工作联系到一起。

从本质上讲数据是存储在计算机磁盘上的一串二进制数，无法被人读取，在数据库软件系统的帮助下，二进制数变成了一条条可读的数据记录。但是数据记录对于税务业务人员来说还是陌生的，因为税务业务人员熟悉的是税收的业务流程，而不是数据库里面干瘪的数据记录。

于是数据仓库诞生了。数据仓库将数据按照具有业务意义的主题组织起来，并依据特定的业务需求对其进行加工，最终以图标的形式将结果展示出来。到这一步，数据已经变成了有业务含义的图表，税务管理人员能够较好地理解和应用。不过因为图表是统计结果，反映的是整个过程和全局的情况，所以对于一线的税务工作人员的意义不大。

针对上述场景，我们认为将税务业务系统中的数据还原为业务流程上的真实涉税事件是帮助税务工作人员理解和利用税务数据的关键。在税务业务系统中除了代码表外的每一张表中的每一条数据背后都是一个涉税事件，比如申报表中的数据代表一次申报行为，纳税人登记表中的数据代表一次新登记或者是信息变更。

我们将数据还原为事件后，还应该进一步对其进行追踪和分析，找出涉税事件间的关联关系，用以辅助对过去涉税事件的纠错和对未来涉税事件的预测。

2. 解决方案

帮助税务人员理解和利用税务数据的关键是涉税事件的追踪分析，其基础是税务数据到涉税事件的还原过程，这需要从税务核心征管系统以及其他各个业务系统数据中根据同一个字段（如纳税人登记证号）进行关联，然后按照记录产生的时间顺序和记录的来源方式形成纳税人涉税事件事实表。在展现形式上采用时间卷轴的方式。

时间卷轴上的每一个亮块就代表一个涉税事件，税务人员能够一目了然地看到某一税源在选定时间范围内的所有涉税事件，并通过单击亮块来查看该涉税事件的详细情况。

对涉税事件的还原和查看只是第一步，接下来还要将在业务流程上有关联的事件标记和关联出来，用来对税务流程中的各个环节进行辅助和提醒。例如，某一税源发生了申报事件，那么接下来一段时间内应该发生缴纳税款事件，如没有找到该事件就推送相关人员知晓。

对涉税事件的关联不应该只局限于税务业务系统内部，还要借助外部数据如互联网数据、公安数据以及银联商务刷卡数据的力量，分析挖掘出更多的信息。目的是做到对已经发

生涉税事件的合规性校验以及对未来将要发生的涉税事件的预测。

具体做法是首先根据涉税事件的人员信息和时间信息去互联网、公安、银联商务数据中匹配同一人同一时间范围内的事件，然后将匹配到的事件作为一个集合，接着对集合中的元素（各个事件）的描述进行精简，只保留谁、什么时间、什么地点、什么动作 4 个要素。同时要根据这 4 个元素准备一个规则引擎，定义事件与事件之间关系的判断标准。

通过规则引擎判断出目标事件和事件之间在逻辑上是顺承的还是违背的，顺承的关系可以用来对未来涉税事件进行预测，违背的关系可以用来对已发生涉税事件的合规性校验。规则引擎的更新由人工智能算法来完成，当然同时还需要设计一个人为干预的反馈机制，用以对规则引擎进行修正。

当集合中的元素都经过规则引擎筛选比对后，将结果通过时间卷轴反映给税务工作人员。以红色亮块表示过去的违规事件，蓝色亮块表示未来的涉税事件。

3. 思维导图

将上述解决方案以图形化的方式展现出来便得到如图 9-9 所示的图形。



图 9-9　对涉税事件的追踪分析

9.5.5 税源画像

1. 场景描述

税源管理一直是税收工作中的重点，管理好税源有利于税款征收、风险控制等工作。当前税源管理采取的是粗放式做法，在政策制定和执行上通常是以大范围为粒度的，缺乏差异性和针对性，导致效果不好。

另一方面，互联网企业尤其是电商行业在用户管理方面依托差异化、个性化的管理策略，在产品设计、商品营销、活动策划等方面取得了很好的效果，为其带来了丰厚的收入。从

他们的经验来看，精细化管理才是大数据时代用户管理的正确方式。

举例来说，精细化管理能够对不同的用户进行不同商品的推荐，采取不同的优惠措施，以此来提高用户的下单率和成交率。税务工作同样如此，只有实现精细化管理，才能让税收工作更上一个台阶。而税源精细化的前提是弄清楚每个税源的特点，然后才能根据税源特点制定差异化的管理措施。

针对上述场景，我们认为对纳税人进行画像是摸清税源特征的最佳办法，从内外部数据入手为税源打上各类个性化标签，从而将税源划分为不同的群体分开进行管理。

2. 解决方案

税源画像本质上是从与税源相关的所有数据中找出税源在各个方面的特征，通过这些来为税源打上相应的标签。一个税源至少拥有一个标签，多个税源可能会拥有同一种标签，这就为我们在税源管理政策制定和执行时提供了依据，能够充分考虑税源的个性化情况去实施，取得的效果当然也会更好。

为此首先要搜集大量的数据，比如国税业务系统中税源的所有涉税数据、税源公安数据、税源银联商务数据、税源互联网数据。其中，国税业务系统中的数据用来分析税源的纳税习惯标签，其余的外部数据则用来分析税源的生产经营状况、经营范围、目标人群等信息。

数据搜集完成后，第一步对国税业务系统中的涉税数据进行分析，统计分析出纳税人的常规缴税时间、缴税渠道、缴税地点、缴税金额等指标，组成纳税人的缴税习惯标签。第二步综合所有的外部数据统计得出纳税人的各项标签，如企业法人信息、注册地址、经营地址、贷款情况、营业额、纳税信用风险评级、资金往来频率、增值税缴纳情况、目标人群等。

最后通过文本解析和规则引擎，分析税源的缴税习惯标签和其他各项标签，辅助制定出对纳税人的精细化管理策略，用以提高税源管理水平。

3. 思维导图

将上述解决方案以图形化的方式展现出来便得到如图 9-10 所示的图形。



图 9-10　税源画像思维导图

9.5.6　纳税遵从指数

1. 场景描述

江苏全省地税系统以风险管理为导向的税源专业化管理模式基本到位，形成了风险管

理计划、风险指标模型建设、风险识别、风险推送、风险应对、风险应对绩效评价的基本管理流程，能基本支撑起新的税源管理模式。但从风险应对结果、纳税人履行征管制度情况等角度监控评价纳税人税法遵从方面还存在缺陷，有必要通过对纳税遵从情况的监控评价来促进税收管理“征、评、管、查”各环节，提高整个税源专业化管理模式运行成效。

2. 解决方案

运用纳税遵从监控评价指标体系，利用主题税源平台提供数据，加工形成纳税遵从监控评价数据，分行业类型、税种税目、风险事项对不同的地区、各个征管环节进行纳税遵从监控评价，输出分类汇总表单，提供分类遵从度评价排序，批量和单户纳税遵从查询，出具纳税遵从监控评价智能报告。

纳税遵从监控评价管理包括如图 9-11 所示 4 个部分。

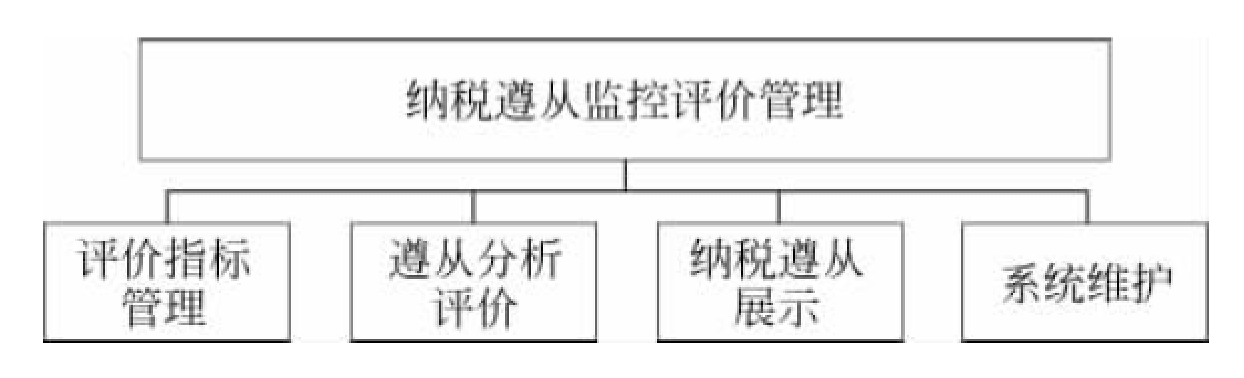


图 9-11　纳税遵从监控评价

（1）评价指标管理。对纳税遵从评价指标进行管理，建立评价指标分类，并按照分类归集各类评价指标，实现评价的动态配置。

（2）遵从分析评价。对已经配置完成的评价指标，运用主题税源平台数据对纳税人进行纳税遵从评价分析，输出纳税遵从结果，并依据规则实现纳税遵从值的计算，按户归集纳税遵从值。建立遵从分析评价的模板，以此判定纳税人的纳税遵从评价度。

（3）纳税遵从展示。按照不同分类组合展示一类、一个地区的纳税遵从情况，按照不同需求查询纳税人纳税遵从情况，并实现按户纳税遵从情况查询。对一个类别纳税人实现动态纳税遵从报告的出具，对单一纳税人出具纳税遵从报告。

（4）系统维护。用于配置系统功能。

根据已经制定的纳税遵从度评价指标体系梳理出来的思维导图如图 9-12 所示。

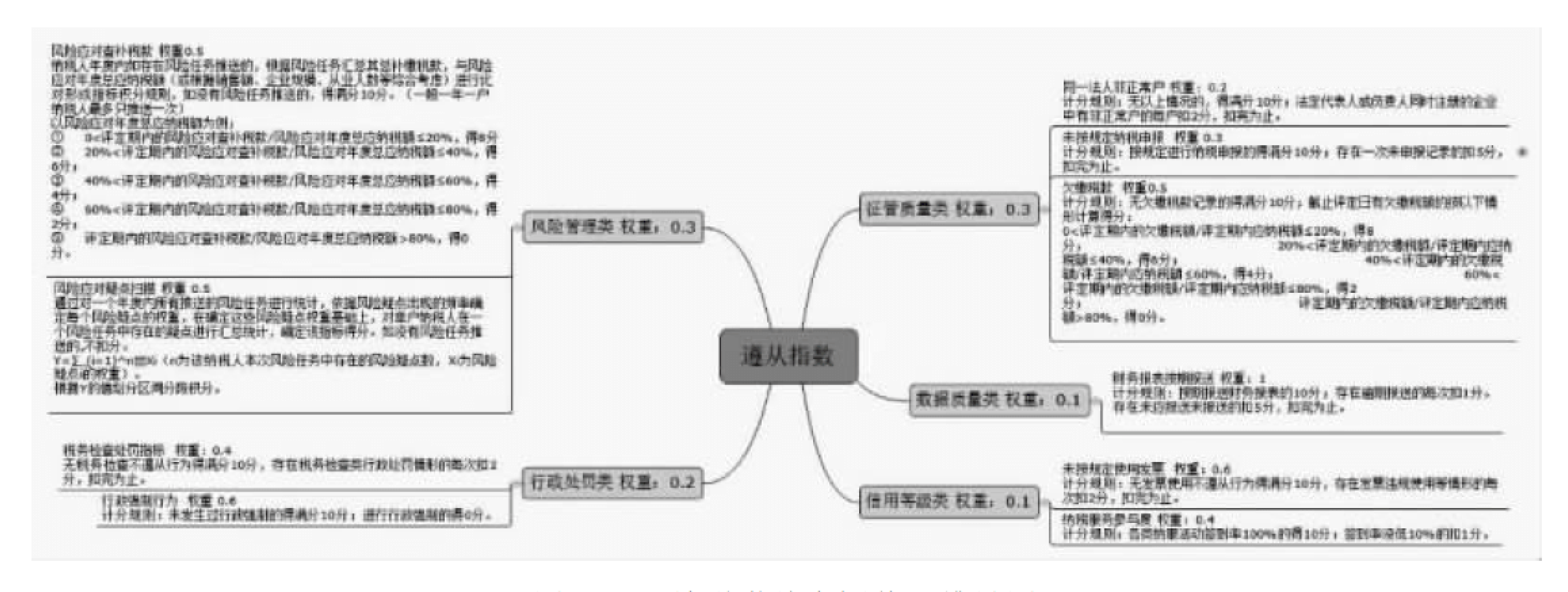


图 9-12　纳税遵从度评价思维导图

可以看出，整个评价指标系统由风险管理类、行政处罚类、信用等级类、数据质量类、征管质量类 5 个大类组成，每个大类细分为若干小类，各个小类又细化到具体的指标，每个指

标有自己的评分规则和权重，通过得分和权重相乘并累加的方式能得到各个小类的得分，再层层汇总，最终就能得到目标的纳税遵从指数。

9.6　税务大数据服务价值

明略数据为税务部门构建的可视化涉税分析平台定位为面向税务部门的数据服务产品。产品充分利用明略底层大数据平台相关技术，数据挖掘建模技术及明略税务行业研究专家对税源管理专业化、风险控制精细化、决策分析智能化的理解，搭建以分析预测为核心的数据应用平台，以帮助税务部门征管工作更有效、更全面、更精细化地展开。

可视化涉税分析平台能够对政府“信息管税”带来以下影响。

1. 成本更加可控，更丰富的数据视角，更敏捷的分析构建

大数据的技术核心在于可扩展性。对用户来说，可扩展性意味着以成本可控的方式逐步进行信息化建设，相对于传统的单机数据仓库构建，基于大数据平台能够以更低成本获得数十倍的数据处理能力提升，堪称技术架构上的飞跃。

传统的一户式视图能够对单户纳税户的所有涉税电子档案信息形成查询，但是数据的类型丰富程度和数据之间关联度都还比较缺乏。可视化涉税分析平台不仅以税源为中心组织数据，还能够大量引入外部各种关系型和非关系型数据，在更大的尺度上进行模型构建，充分建立数据关联，通过撬动大数据的杠杆轻松切换观察视角，敏捷构建业务场景和报表，从数据资产到分析展示不再经历传统冗长的数仓过程。

2. 获得数据之外的洞察

发挥数据的价值，挖掘数据背后的相关性，才能为业务决策带来最大效益的帮助。明略可视化涉税分析平台能够通过分类、聚类、回归等多项算法，发现数据相关性，清晰洞察业务关联信息，更精确地预测走逃税风险。得益于大数据技术带来的性能飞跃，这些在传统架构中复杂耗时的数据训练和建模运行周期从数月能够缩短到数天甚至数小时。在数据时代，掌握数据运用的方法论将成为每一个数据拥有者的首要任务。

3. 打造拥有大数据能力的数据服务层，为进一步数据应用打下基础

所有的大数据项目都不仅是提供一个产品或者完成一个项目，而是让客户获得针对大规模数据的持续服务能力。财税部门既掌握着第一手的纳税人经营状况和财税报告等信息，也拥有个税、车船、土地增值税等经济活动信息，这些信息勾勒出一个行政区域的经济脉络，将有可能成为政策分析、经济状况评估、银行借贷、企业和个人信用评估等经济活动的重要参考凭证。数据的集中治理、一致性的数据口径和大数据平台提供的强大处理能力是一个基础，后续数据和数据之间的相互协同和互补将在更大层面影响和帮助不同的经济活动参与者。

技术的延伸会深刻改变人们生活和工作的方式，归根到底，技术总会在某一个层面极大地提升效率。大数据的落地应用，创新地赋予税务行业成为依靠数据驱动的高效行业，也为这个行业的每个参与角色提供服务和享受服务的可能性。基于数据的生态逐渐完善后，政府部门、纳税企业、自然人都能合理有效地执行自己的权利和义务，而数据会越来越成为维持商业生态的能量载体，高效流转于每一个环节的数据应用中，成为基础服务的原材料和创新应用的助推剂。

另外，随着未来全国所有增值税发票都将纳入网络电子化管理，发票信息也将全面纳入大数据分析，还有税收工作逐渐从大面转向具体的个人，必然会带来税务数据爆炸式的增长，如何管好利用好这些数据为税务工作服务将是税务机关面临的难题。

大数据技术在税务行业的落地实施，会是上述问题的一个解决途径。它能促使税务服务水平大幅提升，管理进一步优化，而且将形成一整套可靠的经济数据，为国家的经济决策提供重要参考。并且透过互联网、移动互联网、物联网形成的大数据环境，还能让税务数据和外部数据联动起来，能更好地做好税务风险评估、税款征收、税款追缴等一系列税务工作，开创税收工作的新局面。



大数据时代的电力服务

国务院总理李克强在十二届全国人大四次会议上指出,“创新驱动发展战略持续推进,互联网与各行业加速融合,新兴产业快速增长。”提出要强化创新引领作用,为发展注入强大动力。强调要持续推动大众创业、万众创新,促进大数据、云计算、物联网广泛应用。2015 年 9 月国务院印发的《关于促进大数据发展的行动纲要》强调,要开发应用好大数据这一基础性战略资源。

随着一系列国家战略与产业发展规划的出台,作为国家战略,大数据已成为国家下一个创新、竞争和发展的前沿,也必然成为企业提升核心竞争力的战略制高点。作为正向能源互联网转型的传统电力行业,大数据及云计算时代的到来将为传统电力行业的发展注入新的活力,传统电力行业有可能产生革命性的变化。

10.1 电力大数据面临挑战

电力大数据具有量大、类型多、速度快等特点,其背后反映的是电网运行方式、电力生产方式及客户消费习惯等信息,这些数据如果能挖掘分析好,就能释放大数据真正的价值。

大数据时代,数据质量的高低、数据管控能力的强弱直接影响了数据分析的准确性和实时性。传统的电力行业数据在可获取的颗粒程度,数据获取的及时性、完整性、一致性等方面的表现均不尽如人意,数据源的唯一性、及时性和准确性急需提升,部分数据尚需手动输入,采集效率和准确度还有所欠缺,行业中企业缺乏完整的数据管控策略、组织以及管控流程。数据共享不畅,数据集成度不高。大数据技术的本质是从关联复杂的数据中挖掘知识,提升数据价值,单一业务、类型的数据即使体量再大,缺乏共享集成,其价值就会大打折扣。目前,电力行业缺乏行业层面的数据模型定义与主数据管理,各单位数据口径不一致。行业中存在较为严重的数据壁垒,业务链条间也尚未实现充分的数据共享,数据重复存储的现象较为突出。

提高数据利用率,便可使企业提高相应的利润。国家电网作为中国电力工业的国有重要骨干企业,早已把发展的目光投向大数据领域,积极探索大数据应用,挖掘大数据商机。电力公司“三集五大”体系和坚强智能电网建设,积累了体量大、类型多、速度快等典型大数据特征的运营数据,具备了推广大数据应用的基础条件。国网智能电网研究院的数据显示,截至 2015 年年底,国家电网公司管理结构化数据 49.75TB,非结构化数据 213TB,营销基础数据 130TB,用电信息采集数据达 43TB,且信息化数据平均每天以 10TB 的速度增长。“数据海量、信息缺乏”是传统能源企业面临的问题。

高度专业化的传统电力企业，并不擅长最大限度地利用数据。随着公司信息化建设不断深入，业务系统产生的数据量呈爆发式增长，电力企业目前面临着大数据带来的海量存储以及部分业务系统面临存储升级成本较高、系统响应速度较慢等挑战。在客户服务方面，公司如何开展客户行为大数据分析，以达到提升客户用电效能，提高服务水平的目的？在电网安全生产方面，如何开展配电网运行效率效益分析，及时发现薄弱环节，优化电网建设投资，提升电网运行水平？在企业经营管理方面，如何开展业务执行效率效益大数据分析，关注运营中"敏感点""发热点"，分析供电服务、资金收支等运营业务合法合规性，防控经营风险、提升运营效率？如何在提升系统访问效率的同时，节约系统存储成本？分析业务系统架构，在可能引起系统访问瓶颈的地方，能否引入大数据技术加以解决？电力大数据面临挑战，需要寻求突破。

10.2 电网运营大数据

10.2.1 电网系统架构现状

电网现有系统在技术架构实现上大多遵循 Java EE 技术体系，按照展现层、服务层、业务层、数据层进行分层设计。展现层根据应用需求采用多种技术实现；服务层则以组件化、动态化为准则，最大限度实现复用，提供 WebService 等方式实现服务的互通互用；业务层实现具体的业务逻辑；数据层则根据数据类型和业务需求的不同，选择不同的方式进行存取，大多数情况下采用商用关系型数据库实现相关功能，对于数据集成方面，则采用 WebService、消息服务完成。

在项目建设上，大量采用 IOE 设备，IBM 小型计算机、Oracle 数据库、EMC 存储组件，该技术架构在传统领域中得到了很好的验证，针对传统的业务需求，如 ERP、CRM 等均能满足需求。但随着系统建设从功能建设到决策支持建设的转变，业务系统的服务对象，从满足部门、网省公司需求，到满足国网公司需求，直至服务于国网公司的整体战略需求，对业务数据的深度挖掘应用，以及海量数据的分析处理，成为系统的发展方向。在此背景下，大数据的出现将成为必然。例如，为了保证国网公司的保障电力供应战略，对特定地区的智能电表实时采集数据进行分析，采集数据为 TB 级，同时要求在此 TB 级数据上快速计算出相应的指标值。

基于 IOE 构建的系统架构在面对大数据情况下，存在以下三个弊端。

（1）商业软件架构，为满足大数据的需求，其建设成本急剧增长；

（2）商业系统大多属于大而全的系统，在大数据情况下，大而全的系统，反而暴露出无法很好解决极限性能等问题，从而成为整体方案的瓶颈；

（3）大多数产品不具备横向扩展能力，即无法通过增加物理节点的方式获得性能提升，导致该类产品必然遭遇性能瓶颈。

10.2.2 高性能架构设计

电力大数据具有量大、类型多、速度快等特点，其背后反映的是电网运行方式、电力生产方式及客户消费习惯等信息，这些数据如果能挖掘分析好，就能释放大数据真正的价值。

针对电网，我们提出了需求分析指标法，提出对需求性能瓶颈的考量方式。我们将分析

指标分为存储指标和计算指标两类,标注不同应用场景的区别,如图 10-1 所示。



图 10-1　分析指标

1. 数据存储指标

数据存储指标是对业务需求中数据本身结构和操作方式的抽象概括,从而明确该类数据在存储时的性能问题。数据类型主要分为元数据、主数据、历史数据、交易数据、机器数据、Web 和社交媒体。

1) 元数据

元数据(Metadata)是指描述数据的数据,主要是描述数据属性的信息,用来支持如指示存储位置、历史数据、资源查找、文件记录等功能。元数据算是一种电子式目录,为了达到编制目录的目的,必须描述并收藏数据的内容或特色,进而达成协助数据检索的目的。

该类型数据的特点往往在于数据量小,结构简单,需支持高并发访问,但数据关键,一旦丢失将导致其他系统无法完成。

2) 主数据

主数据是指在整个企业范围内各个系统(操作/事务型应用系统以及分析型系统)间要共享的数据,比如,与客户、供应商、账户以及组织单位相关的数据。但需要注意的是,主数据不是企业内所有的业务数据,只是有必要在各个系统间共享的数据才是主数据,比如大部分的交易数据、账单数据等都不是主数据。而像描述核心业务实体的数据,如客户、供应商、账户、组织单位、员工、合作伙伴、位置信息等都是主数据。主数据是企业内能够跨业务重复使用的高价值的数据。这些主数据在进行主数据管理之前经常存在于多个异构或同构的系统中。

该类型数据的特点是数据量中等,关系复杂,新增操作多于修改操作,需支持复杂的关系查询。

3) 交易数据

交易数据是产生于交易活动中的描述性数据,由于交易活动复杂、相互关联,因此该类型数据经常伴随有时间维度、可量化的数字和一个或者多个关联实体。例如,订票业务、话费业务等都产生大量交易数据。该类型数据是对交易活动的描述,一般数据量可预估。

在高性能瓶颈上,该类型数据一般会面临以下两大问题。

一是并发性。由于该类数据往往会遇到针对某一资源争抢,如抢票操作等,因此该类业

务系统会产生瞬值访问高峰。

二是事务性。数据操作严格遵守顺序，必须保证操作的原子性，这对于分布式操作提出了极大的挑战。

4）历史数据

历史数据在本处特指历史行为产生的数据，该类数据不会进行更改，但量大，如航班历史飞行计划、用电历史数据等。前三类数据均可变为本类数据，但在变化为历史数据后，是不可修改的。

该类型数据特点是数据量大，不会进行修改，操作上更多以总体查询为主，对于个体明细查询要求低。

5）机器数据

机器数据指由计算机进程、应用或者其他机器在人力没有参与的情况下自动产生的数据。该类型数据强调了非人力参与，是对人类活动的观察而不是对人类的选择进行的。

机器数据的显著特征在于数据量大，结构单一，未来增长较快。该类型数据包括 Web 服务器日志、智能电表数据等。

在高性能瓶颈上，该类型数据一般会面临以下两大问题。

一是写入性能。由于该类数据产生速度极快，数据量大，因此对数据写入操作提出了极高的要求，例如智能电表数据，一个省每秒的数据就可以达到 GB 级别，传统数据库难以应对如此巨大的数据写入，同时，如果没有进行良好的读写分离，数据库必然会面临崩溃问题。

二是查询性能。机器数据意味着海量的数据存储，如何快速地在如此巨大的数据中完成查询操作，同样是一大挑战。

6）Web 和社交媒体数据

Web 和社交媒体数据指在 Web 2.0、智能终端普及情况下产生的新的一类数据。该类数据产生于 Web 应用以及社交媒体如微博、微信等。

其显著特征在于，数据格式各异，如图片、文字、音频等，同时数据直接反映了用户群体的直接需求，具有极大的利用价值。

在高性能瓶颈上，该类型数据主要设计考虑点在于异构文件存储和关联查询。

2. 数据计算指标

数据计算指标是对业务需求中数据交互、计算等流程的抽象概括，核心指标为计算时间以及计算复杂度。

1）计算时间

实时计算：实时计算是指数据计算要求在毫秒级别完成响应的计算业务。

准实时计算：准实时计算是指计算要求在秒级别完成响应的计算业务。

离线计算：离线计算是指计算要求在分钟级别及以上完成响应的计算业务。

2）计算复杂度

统计型应用：统计型应用是指以统计为主的业务应用，是基于现有数据进行的统计计算，无须复杂算法介入。

数据挖掘型应用：数据挖掘型应用是对统计型应用的深化，需要采用多种数据挖掘算法进行计算的应用，该类型应用比统计型应用具有更大的计算量及更复杂的逻辑关系。

查询报表型应用：查询报表型应用是指对精确数值查询及统计分析结果展现的应用，

相比统计型应用、数据挖掘应用，该类型应用关注于数据的个体，需要对个体实现精确展现。

其他：除去上述三类应用的其他类型应用，如实时ETL操作等，该类型应用则需要根据实际需求进行合理的定制开发。

10.2.3　技术选型

通过需求分析指标法，能够明确不同业务需求下适应的技术，解决大数据情况下，可选用技术过多，无法快速定位技术应用场景的问题，为高性能架构设计做出指导。

1. 存储类性能需求技术选型

数据存储类应用技术选型准则如表10-1所示。

表10-1　数据存储类应用技术选型准则

数据类型	技术选型		
	1MB～1GB	1GB～100GB	100GB以上
元数据	关系型数据库	内存数据库	HBase
主数据	关系型数据库	分布式缓存技术＋关系型数据库	HBase
历史数据	关系型数据库	HBase	HBase
交易数据	关系型数据库	NewSQL	HBase
机器数据	HBase	HBase	HBase/Hive
Web和社交媒体	HBase	HBase	HBase

在元数据业务应用中，由于元数据的重要性及并发访问量高的特性，同时数据关系简单，因此在数据量小于1GB的情况下，采用现有关系型数据库即可满足要求。当数据量大于1GB时，系统的并发访问瓶颈将凸现出来，必须采用内存数据库的技术，更多地将数据加载在内存中保证高并发读写能力。由于元数据的丢失将导致大量系统的不可用，因此该类型应用不适合使用纯内存数据库，如Redis等，推荐使用LevelDB等半持久化内存数据库。当数据量达到TB级别以上时，则HBase成为首选，由于HBase本身的高吞吐量、响应准实时特性，从而保证了满足系统要求。

主数据应用中，由于数据大量的为企业名录、员工信息等，该类型数据的特征是一次导入后修改次数较少，修改频率低，属于典型的读多写少的应用，因此在架构设计上应该优先考虑读取性能，同样地在小数据量情况下，关系型数据库依然可以满足需求。当数据量上升到GB级别时，则推荐使用分布式缓存技术和关系型数据库的组合，由于数据读多写少，因此通过分布式缓存技术对数据库信息进行全面加载可以保证数据访问的高命中率，同时对现有架构也能很好地结合。而数据量上升到100GB级别后，则建议采用HBase进行实施，由于数据量的大幅度增长，分布式缓存技术加关系型数据的方案难以满足线性扩展的能力，只有通过系统的改造，引入HBase满足性能需求。

历史数据应用与主数据应用类似，都属于读多写少应用，但是在数据粒度上，历史数据应用需要支持大量的聚合类型操作，如求和，而主数据应用则是关注于个体值详细结果。由于该处的不同，历史数据推荐采用Hive作为数据仓库，不仅易于与离线计算框架整合，同时能够将结构化、非结构化数据进行统一存储管理，利于数据应用的展开。

交易数据应用在高性能领域中是难以解决的问题，因为交易数据的事务特性导致高性

能架构理论中的分布式、并行化难以实施，同时复杂 join 操作也将对系统整体性能产生极大影响，因此该类型数据应用与业务密切相关。对于 GB 级别以上 100GB 以下交易数据应用，推荐采用 NewSQL 技术予以解决，如 MySQL Cluster、VoltDB 等。该类型技术均有明显的业务系统定制开发特性，例如 MySQL Cluster 分区分片操作必须对原有业务系统进行详细的理解，而 VoltDB 则需要对数据表结构进行分区设计，保证整体性能。但由于在大多数业务系统中，交易类型的数据一般均小于 GB 级别，因此该类型数据使用关系型数据库即可完成，相关的统计分析应用则通过 ETL 转入到数据仓库中，实现 OLTP 和 OLAP 的分离。

机器数据应用往往会与后期分析挖掘相结合，由于机器数据结构简单、产生速度快、数据增长量快，同时不存在数据更新问题，因此在建设初期就可以采用 HBase 进行数据统一存储。HBase 具备在 100GB 以内数据量的查询、统计、分析能力，特别针对结构简单的机器数据 HBase 的 KV 存储模式易于实现。当数据上升到 100GB 以上时，根据访问时间响应和查询粒度可以分别走两条技术方案。HBase 提供更好的数据查询响应时间及细粒度的查询结果展示，Hive 则提供更好的范围统计能力，但退化为离线处理。

Web 和社交媒体数据常常应用在用户价值分析、企业声誉检测等业务场景中，由于数据的多样性、数据量的不可控性等，HBase 成为该类型数据存储首选。

综上，当数据量在 100GB 以内时，可以选择多样的解决方案，但是当数据量大于 100GB 时，HBase/Hive 是最为理想的解决方案。

2. 计算类性能需求技术选型

数据计算类应用技术选型准则如表 10-2 所示。

表 10-2 数据计算类应用技术选型准则

时间维度	处理目标	技术选型			
		1MB～10GB	10GB～100GB	100GB～1TB	1TB 以上
实时计算	统计型应用	Storm	Storm/HBase	—	—
	数据挖掘应用	Spark＋Mlib	—	—	—
	查询报表型应用	关系型数据库	关系型数据库	—	—
准实时计算	统计型应用	—	—	Impala	—
	数据挖掘应用	—	Spark＋Mlib	Spark＋Mlib	—
	查询报表型应用	—	—	HBase	HBase
离线计算	统计型应用	—	—	—	Hive
	数据挖掘应用	—	—	—	Hadoop＋Mahout
	查询报表型应用	—	—	—	—

为了达到实时计算目标，统计型应用在数据量在 10GB 以下时，首选方案是采用流处理技术、关系型数据库进行解决，例如 Storm 即可完成大量的工作。随着数据量增多，Storm 对资源的占用将对业务系统产生影响，因此当数据量过大后，采用 HBase 不仅可以解决系统资源占用问题，还可以简化平台组件维护成本。但是涉及数据挖掘应用时，由于数据挖掘算法中存在大量的迭代运算，Hadoop 平台是无法满足性能需求的，数据量在 10GB 以内时采用 Spark 加上 MLib 的方式可以满足性能要求。对于 10GB～100GB 的实时查询报表场

景，Storm 和 HBase 则是最适合的解决方案。

针对准实时计算目标，由于响应在秒级，针对 100GB～1TB 级别的数据的统计分析型应用采用 Impala 处理，不仅可以支持 SQL，同时具有良好的扩展性，而在数据挖掘应用方向依然首选 Spark＋MLib 的方案，但该方案无法应付 TB 级以上数据的准实时计算。HBase 则是大数据量下的处理利器，针对 100GB 以上的数据能最大发挥其优势。

离线计算性能要求只出现于 1TB 数据以上的统计型和数据挖掘场景，针对如此大量的数据，采用 Hive 和 Hadoop＋Mahout 的解决方案是扩展性最好的实施方案，如图 10-2 所示。

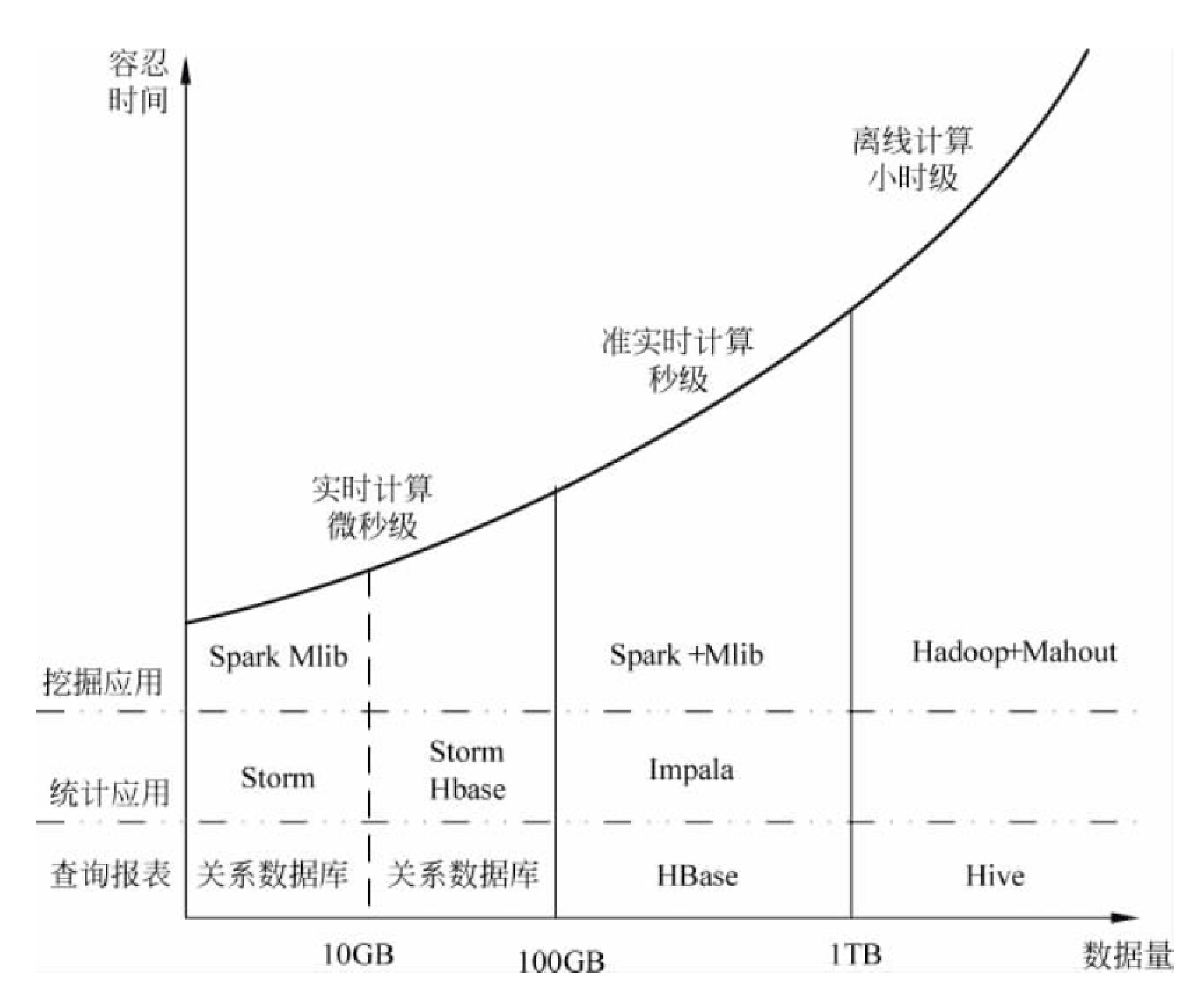


图 10-2　各类技术的容忍时间

综上，针对 100GB 以下数据量处理时，查询报表型应用采用关系型数据库是最为合理的解决方案，不仅可以合理地沿用原有架构，同时通过合理的优化，依然可以达到高性能要求。考虑到数据的增长率，HBase 作为分布式数据库应该及时介入到系统架构中，与关系型数据库进行互补。

统计型应用，在实时响应条件下，采用流处理技术将批处理转为流式处理，提高计算的实时性，能够极大地改变现有架构下对异常数据监控、监测的能力。但随着数据量的增多，实时计算不再可能，通过 HBase 实现准实时方案，是更好的方式。

数据挖掘应用由于其计算复杂，具有大量迭代运算需求，在 100GB 场景下采用 Spark 框架，充分利用其迭代优化特性，能够达到实时、准实时的计算响应时间。数据增至 TB 级别后，Spark 技术方案依然可以满足需求，但考虑到此数据量情况下的挖掘对时间的敏感度进一步降低，但对单位计算成本要求更为苛刻，所以基于内存的 Spark 框架在单位计算成本上更为昂贵，采用 Hadoop＋Mahout 的数据挖掘解决方案，可以获得更为廉价的整体效应。

10.2.4 高性能架构设计实践

1. 背景介绍

运营监测系统中的资金收支管理主要针对营销的售电数据、财务的资金变动、银行账户等数据进行监控，主要包括资金流入、资金存量、资金流出以及应收票据 4 大功能。以资金流入(该功能的数据体量最大，其他的功能类似)为例：需要汇总各省公司的营销、财务数据，通过计算显示每小时、当日、当月的汇总数据(总部、省、地市)以及相关的明细数据。现有架构如图 10-3 所示。

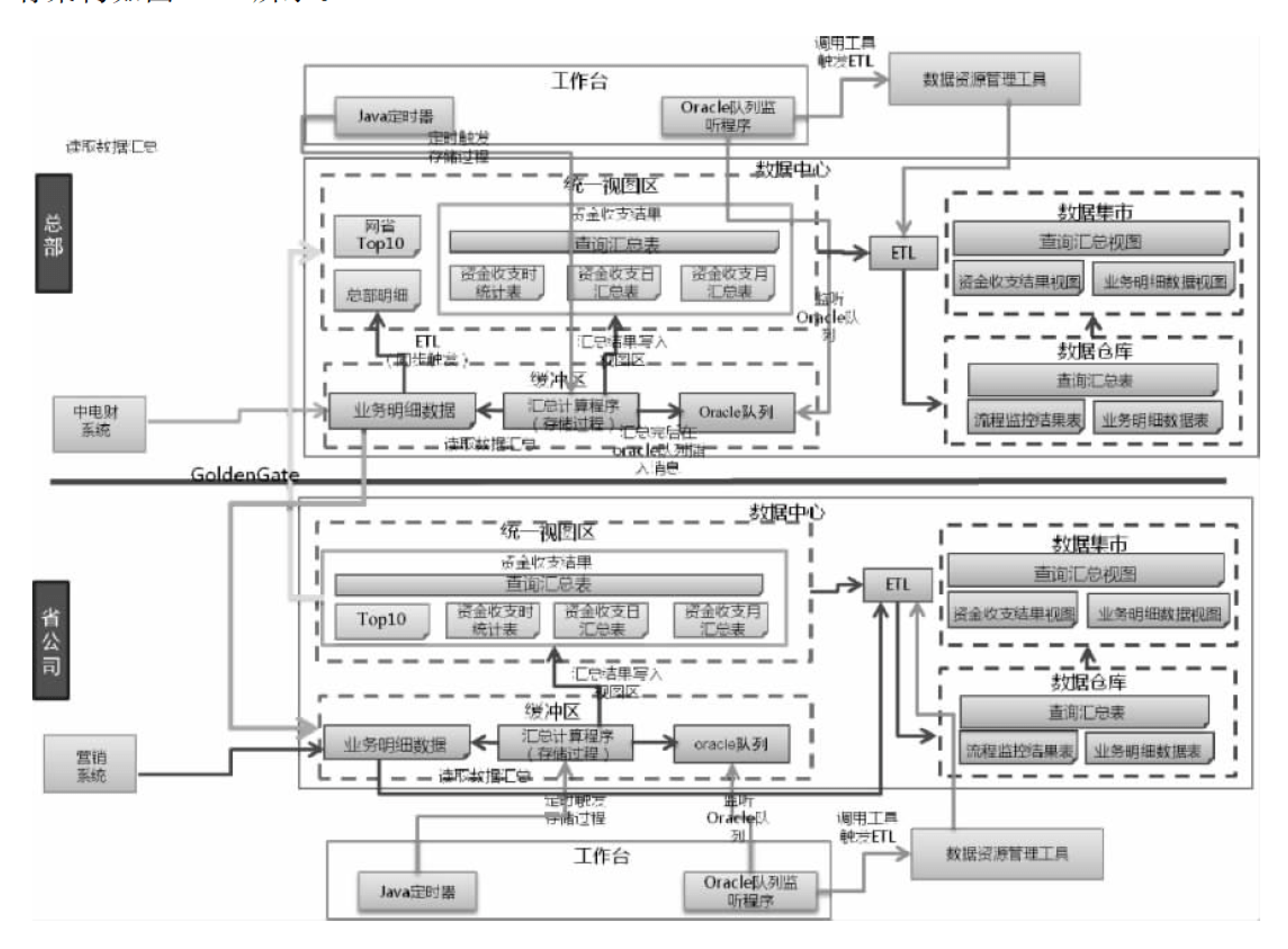


图 10-3　电网现有架构

根据现有方案设计，主要存在以下问题。

(1) 根据设计要求，需要每 5min 进行一次汇总计算，从计算性能和系统整体稳定性等方面考虑，通过定时器触发存储过程的方式难以满足实际需要。

(2) 此方案涉及的数据规模，只考虑 110kV 以上的用户(约 290 万)和全部的公司账户数据，在将来会考虑监测全部的三亿用电用户，数据量将会暴涨，因此在存储和计算上提出了更高的要求。

(3) 系统内部各个模块主要采用 ETL 方式进行数据流转，数据流转效率较低，难以满足实效性较高的需求。

(4) 系统整体横向扩展能力不高，且灵活性不高。

10.2.4 高性能架构设计实践

1. 背景介绍

运营监测系统中的资金收支管理主要针对营销的售电数据、财务的资金变动、银行账户等数据进行监控，主要包括资金流入、资金存量、资金流出以及应收票据4大功能。以资金流入(该功能的数据体量最大，其他的功能类似)为例：需要汇总各省公司的营销、财务数据，通过计算显示每小时、当日、当月的汇总数据(总部、省、地市)以及相关的明细数据。现有架构如图10-3所示。

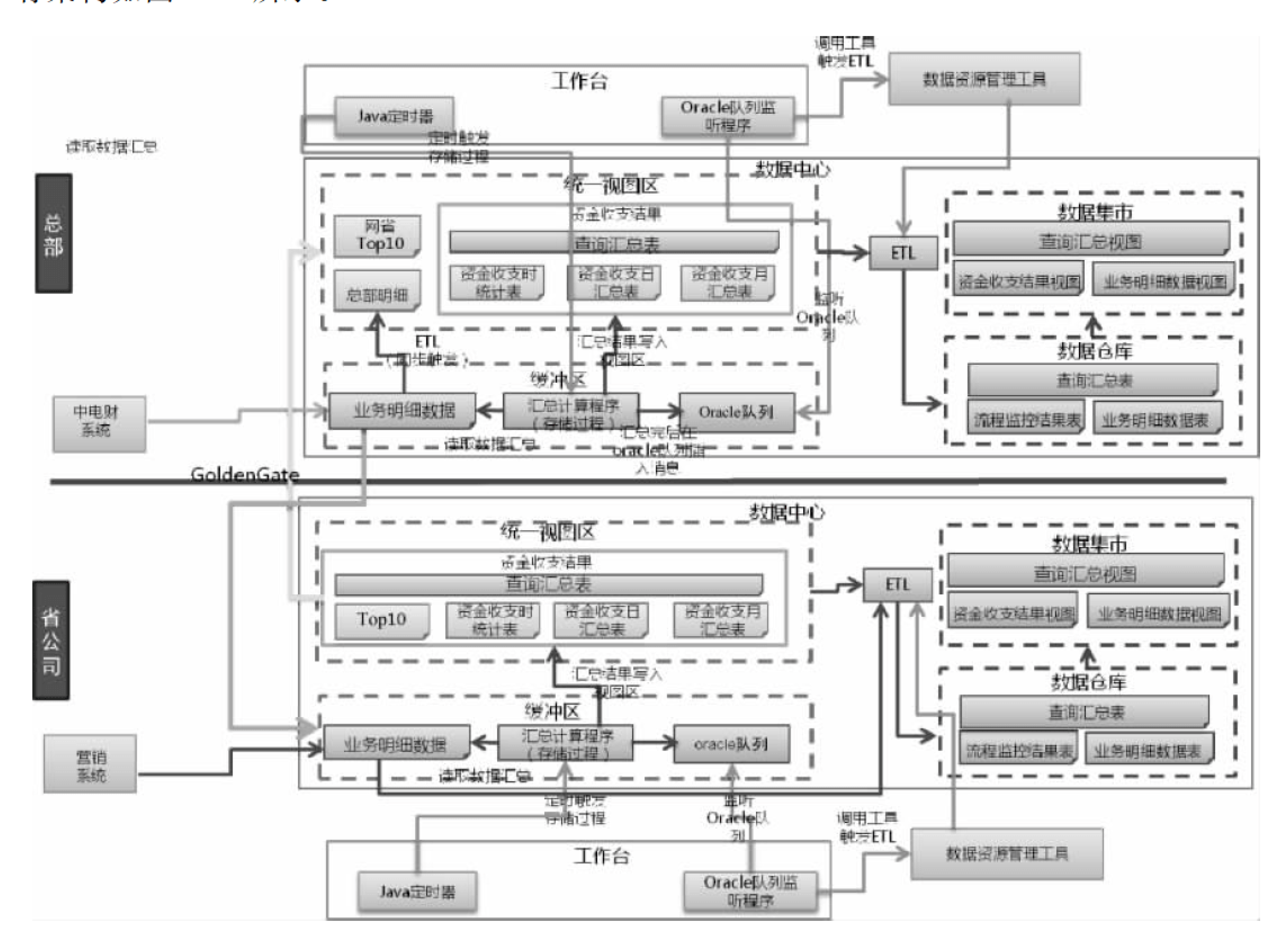


图10-3 电网现有架构

根据现有方案设计，主要存在以下问题。

(1) 根据设计要求，需要每5min进行一次汇总计算，从计算性能和系统整体稳定性等方面考虑，通过定时器触发存储过程的方式难以满足实际需要。

(2) 此方案涉及的数据规模，只考虑110kV以上的用户(约290万)和全部的公司账户数据，在将来会考虑监测全部的三亿用电用户，数据量将会暴涨，因此在存储和计算上提出了更高的要求。

(3) 系统内部各个模块主要采用ETL方式进行数据流转，数据流转效率较低，难以满足实效性较高的需求。

(4) 系统整体横向扩展能力不高，且灵活性不高。

2. 业务分析

运营监测的数据存储分析：该业务数据来源为营销数据，数据类型为 110kV 以上的用户（约 290 万）和全部的公司账户数据，因此属于主数据类型和交易数据，但由于数据进入系统后不再改变，则可以统一视为历史数据。同时又由于未来数据将从现在 100kV 以上用户扩展到监测全部的三亿用电用户，因此数据对存储的性能要求在于 TB 级别的数据存储及明细数据查询，对照存储类性能需求技术选型准则，采用 HBase 进行数据存储是极其适合的。

运营监测的业务计算性能分析：该业务完成每小时、当日、当月的汇总数据（总部、省、地市）的指标计算，并及时发觉异常数据。因此该业务场景属于实时计算下的统计分析应用。根据计算类性能需求技术选型准则，采用 Storm 技术满足实际需求。

3. 架构设计

在完成技术选型后，确定系统采用 HBase 集成 Storm 的方案进行架构设计，得出方案架构如图 10-4 所示。

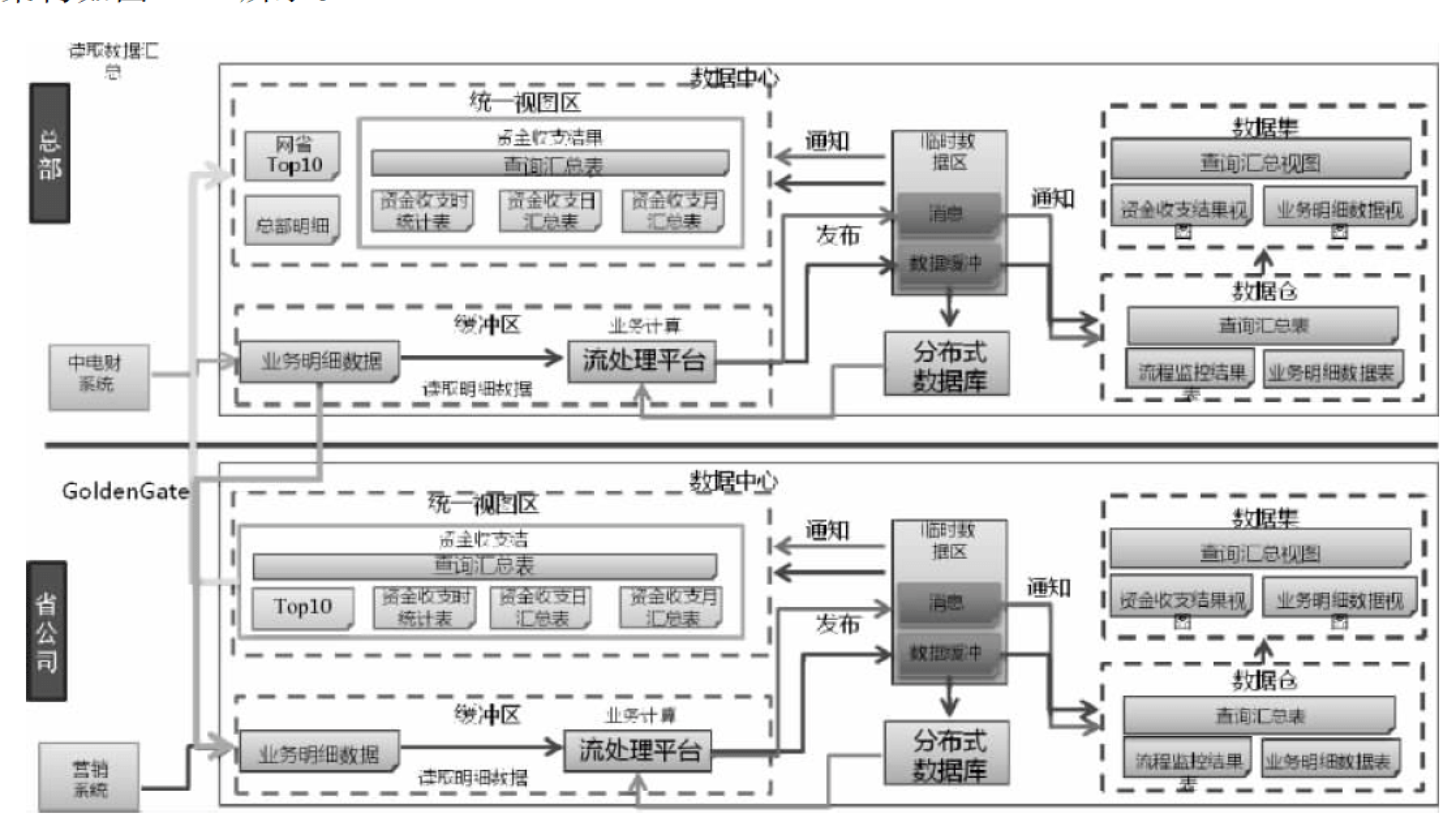


图 10-4　架构示意图

逻辑部署示意图如图 10-5 所示。

在业务明细数据同步到缓冲区后，流处理平台通过任务定时抽取同步过来的数据，生成流处理平台的输入流。在未来也可以通过消息驱动的方式，在同步数据的同时发送消息给流处理平台，实时抽取业务明细等原始数据。

在流处理平台根据业务规则，实时完成收支计算、Top10 统计、原始数据规约化处理等业务功能。处理完的结果会输出到临时数据区并发布数据更新消息。同时流处理系统可根据情况，分别从 HBase（主要存放之前的流处理结果，及一些必要的原始或中间数据）和缓冲区（实时采集数据）获得相应数据（如修正、补偿），并进行相应分析计算。分析计算的结果存放在 HBase 中，供后续的分析使用，或是为数据仓库提供分析结果。

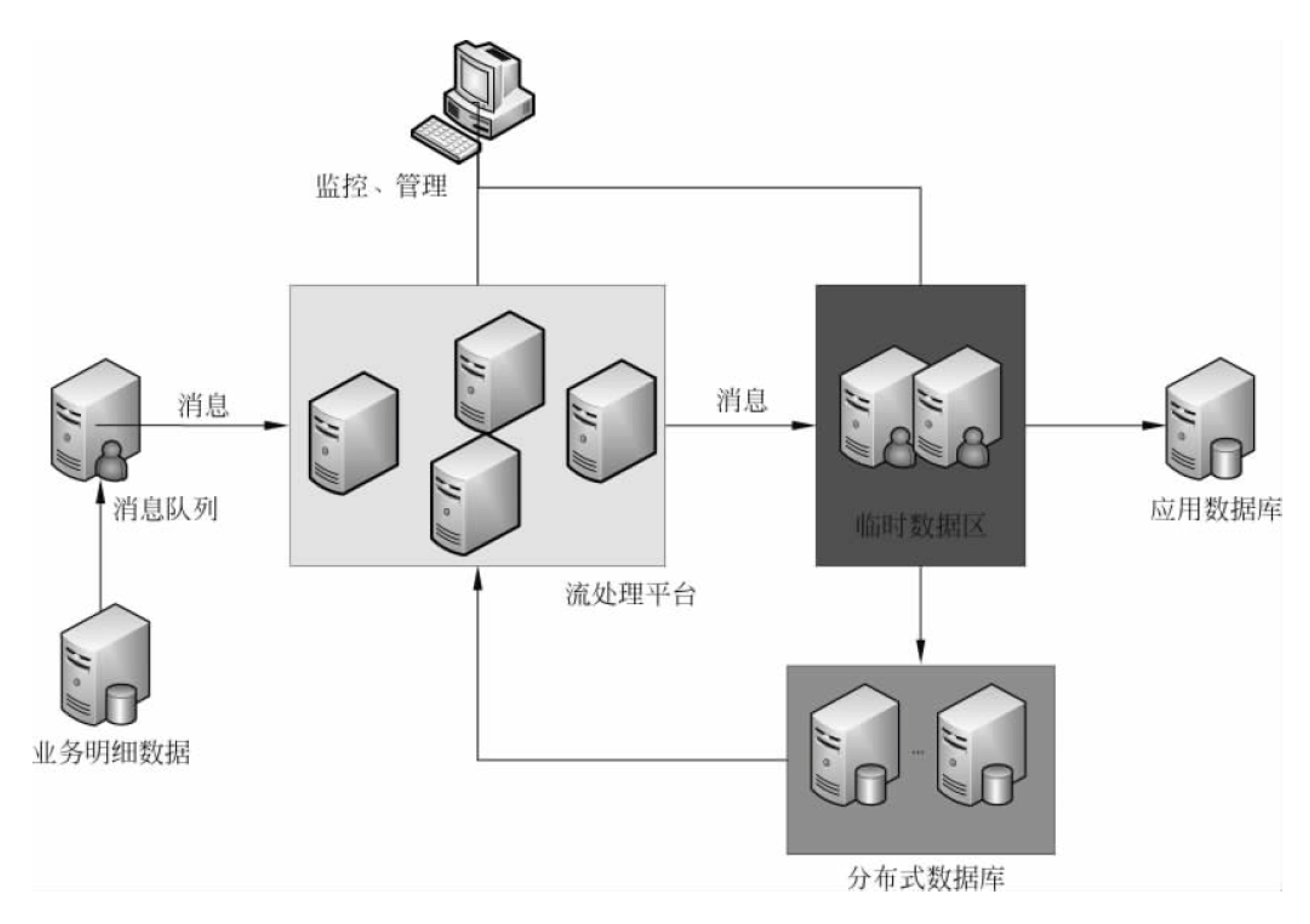


图 10-5　逻辑部署示意图

统一视图区、数据仓库等系统，需要根据实时分析计算的结果做进一步处理，它们会在临时数据区订阅相应的主题。这些系统接收到相应的消息后，会从临时数据区的缓冲区中抽取所需要的数据。

缓存区内的数据会根据预先设定的策略同步到 HBase 中。由于 HBase 集高吞吐、大开发、复杂计算、高扩展及海量数据存储于一身，在将来进行升级时，也可拓展到其他模块中作为海量存储及计算分析服务平台。

4. 高性能架构整体评价

通过业务分析确定了高性能架构存在的性能瓶颈，从而在架构初期选择适当的技术，避免了未来性能瓶颈的出现。

运营监测系统采用了上述方案后，具备了以下 4 大优点。

(1) 用事件触发机制替代了 ETL 轮询方式，用事件流取代批量数据导入，从而增加运监系统实时数据分析的效率，由于去除了过多的 ETL，保证了数据的及时性的问题。同时，充分利用流计算的高扩展特性，提高系统各个子模块数据流转的效率，实现多业务系统海量数据的实时监测的要求。

(2) 架构以分布式技术为基础，流处理及消息队列都采用分布式框架实施，从而保证系统具有高可扩展性、高可靠性、高负载性。

(3) 通过临时数据区将实时计算层和数据消费者之间解耦，增加了系统的灵活性。

(4) 引入分布式数据库，提供数据库的双轨运行，在兼容现有系统下，逐步实现整体架构转向分布式数据库方向，从而避免未来系统一级部署情况下超大规模海量数据的存储、处理和访问带来的种种难题，提供了包括高可用、高扩展、高性能在内的诸多特性。

目前，电网发展已经进入坚强智能电网发展阶段，全球能源互联网正是坚强智能电网发展的高级阶段，具有网架坚强、广泛互联、高度智能、开放互动的特征。而这些特征与大数据都息息相关。

依托大数据，能源管理智能化已经成为新的大趋势。一方面，企业可以利用大数据分析电或其他能源的购买量、可分析预测能源消费和使用，管理能源用户、提高能源利用效率，降低能源成本等。另一方面，以大数据为核心的智能电网的发展，涉及从发电到用户的整个能源转换过程和电力输送链，与智能电网相关的，如智能电网基础技术、大规模新能源发电及并网技术、智能输电网技术、智能配电网技术及智能用电技术等，将优化能源的生产方式、利用方式以及消费方式，如清洁能源、电动汽车的发展和利用等，催生新的经济模式，这都是未来电网的发展方向。

在能源互联网＋新电改的背景下，在大数据＋云计算的新时代，依托电力大数据的电网将迈进全新的时代。

10.3　电网用户行为分析

某区域的用电行为分析是电网企业应用大数据分析技术的一个具有代表性的重要案例。此区域用电行为分析是通过对区域主网负荷和用户负荷有关数据挖掘分析，从中发现宏观主网层面和微观用户层面的用电规律和特点，并将这些规律与电力生产、调度和营销策略等相结合，促进企业运营效率提升，使企业牢牢占据电力生态系统的核心位置。

用电行为分析工作需要持续稳定的智力支撑和高素质的队伍支撑。一是成立业务专家队伍，为分析工作提供业务专业指导和支持；二是建立部门级的联席会议制度，对有关分析报告进行部门会审；三是以外部的专业的大数据分析公司作为支撑机构，合力锻造完成这个大数据分析任务。在这个组织机构下主要完成下述两方面的工作。

在宏观主网层面，主网运行期间出现的负荷种类多种多样，如居民用电负荷、工业负荷、商业负荷等。随着经济的快速发展和人民生活水平的不断提高，由于人口增加、企业发展、产业增多、工业生产等多方面因素所致，这些负荷的持续增加对主网负荷产生较大影响。为维持电力系统的稳定、安全、可靠运行，需要避免主网在超负荷状态下运行，尽量保持在一个合理的水平区间下运行。

在微观用户层面，单个用户的用电数据是这个用户关于电力这个特殊产品下的消费和行为数据，隐藏着该用户的用电习惯。对这些用电数据进行挖掘并研究用户类型，可以帮助电网企业了解用户的个性化、差异化服务需求，从而使电网公司进一步拓展服务的深度和广度，为电力需求侧响应提供数据支撑。

用电行为分析和一系列配套行动工作，有助于电网企业在竞争中脱颖而出，成为最终赢家。这种独具创新的方法一方面能实现错峰用电，更好地平衡各类用电负荷的实际需求，避免出现各个地区之间的电能供应失衡；另一方面能增强电网企业和用户之间的互动，充分提高居民和非居民用电的节能意识。用电行为分析对电网企业意义重大。本章重点对某区域的用电行为分析的流程、技术和经验进行详细阐述。

10.3.1　分析目标及原则

随着售电市场的放开，消费者越来越关注智能用电领域的进步，以及各种突破常规供电

服务的技术创新，电力消费市场正呈现出千差万别的发展面貌和格局。国务院“新电改”政策及消费者需求不断升级对电网企业产生的压力不言而喻，从而刺激电力供应商更关注消费者的参与度和满意度。电网企业必须在关注并努力加强已有业务运营工作的同时，想方设法地继续构建未来核心竞争力，从而确保在未来开展“常规业务”。

随着电网公司营销信息化工作的快速推动，客户用电基础信息不断完善，用电信息采集范围和采集成功率逐步提高，营销业务在线数据应用能力显著增强，为应用大数据挖掘技术，更准确更有效地挖掘客户用电特征和用电价值奠定了数据基础。

基于大数据技术的客户用电行为分析更注重对客户用电价值挖掘，可以实现对客户用电行为定量分析，提高客户行为定位的准确度，为更有效地开展客户服务，提高客户满意度、降低电网运营风险提供决策参考。

国外对于用电行为分析主要集中在对智能家居的用电进行分析，较国内起步和发展早，需要采集大量的智能家居的用电详细信息，不适用现阶段国内的实际情况。

国内对于用电行为分析主要是传统数据量下的分析方法，主要包括专家经验法、统计分析方法、无监督学习法等。这些分析方法可在一定程度上对用电行为进行预测，但在大数据情况下应用效果并不是很理想。

如专家经验法，随着居民用电信息采集量上升，进行用电行为分析耗时耗力；如统计分析法，依赖于大量的家电设备自身的信息，采集这些数据难度较大，不符合目前的现状；如无监督学习，在传统数据量下具备一定指导价值，但在海量数据下算法运行的性能得不到保障。

同时，这些方法在分析用电行为时，将主网的用电特征和居民的用电特征分割开看，在落地指导实践时，忽视了整体规模效益，侧重个体分析，容易造成理论和实践之间的无缝连接效果打折扣。

针对上述缺陷，本案例以某区域为试点，依托宏观层面和微观层面的两个用电行为分析模型，以某区域的主网和该区域用电客户为分析对象，通过用电信息明细数据开展相关分析，重点对用电行为分析的流程、技术和经验进行详细阐述。

1. 工作目标

根据电网公司营销部的职责定位，用电行为分析业务的总工作目标是：充分发挥营销业务信息系统功能，实时汇总分析各类业务和客户信息数据，为公司经营决策提供有力支撑，为提高管理效率和经济效益提供保证。具体工作目标包含下述两个方面。

（1）宏观层面用电行为分析。首先对某区域的宏观主网负荷信息进行特征分析，获取不同时期负荷的曲线特点；然后，在这些群体下，进一步分别针对居民或非居民进行对应的用电情况分析。最后，将主网负荷情况和居民、非居民的情况匹配起来考虑。

（2）微观层面用电行为分析。仅考虑每个用户自身的海量用电数据。分析单用户在时间维度下，不同时期的差异化的用电负荷特征。

本案例在客户群分类方面，在系统中以区域、行业、电价单一属性定性分类的基础上，通过数据分析方法，与客户详细的用电行为特征信息实现有效挂钩，从而更有针对性地提出差异化服务策略。

2. 工作原则

准确把握电网公司“你用电，我用心”的战略目标，以提升公司整体运营效率和效益

为目标，以负荷采集和监测为基础，围绕负荷采集和监测发现的异动和问题，从公司整体运营的高度，以先进管理理论和分析方法为指导，开展跨专业、跨部门的用电行为分析，揭示问题成因、影响及风险，提出对策建议，为公司经营决策提供有力支撑。遵循以下工作原则。

1）全局性原则

用电行为分析工作，应站在公司全局的高度，重点开展跨专业、跨部门、围绕用户的深度分析，反映公司整体运营的效率和效益。

2）客观性原则

用电行为分析工作，既依托公司营销部相关业务部门的专业分析，又独立于专业分析，确保分析结果的客观性。

3）科学性原则

用电行为分析工作，应以先进的管理理论为指导，运用科学的分析方法，构建科学的分析模型，确保分析工作的科学性。

4）有效性原则

通过用电行为分析工作，揭示问题和异动的成因、影响及风险，提出改进建议和措施，为公司决策提供有力支撑，提升管理效率和经济效益，确保分析工作的有效性。

5）创新性原则

用电行为分析工作，应结合外部数据和信息，应用先进分析技术，创新分析思路和分析方法，提升分析质量。

10.3.2　用电行为分析总体架构

对于电网用户的行为分析主要从宏观和微观两个层面进行。

一是宏观层面用电行为分析。通过总结主网负荷下不同客户群负荷特性规律，可以有效识别具体移峰填谷潜力大客户，为公司开展有序用电管理、电网规划建设提供参考。实现思路是在研究电力业务的基础上，根据某区域一年内的主网用电负荷信息，对主网的负荷进行不同日期的分群。进一步根据这些主网负荷群体特点，结合用户的客户档案信息、用电行为数据等，利用基于 Map Reduce 下的 K-means 的聚类方法进行二次聚类，完成海量数据下的用户用电行为分析，以更进一步提高电网需求侧能效管理水平，为公司决策提供更有针对性的参考依据。

二是微观层面用电行为分析。基于通过收集、归类和定义不同属性和行为特征的单一客户，分析单一客户在不同时间的差别用电服务需求，为公司有针对性地开展客户服务工作，提升客户服务效率提供参考。实现思路是根据居民或非居民的自身负荷曲线在一年内的特征，采用基于 Map Reduce 下的 K-means 分别对不同的时期单用户负荷进行聚类，获得居民及非居民负荷曲线在不同日期下的分群。这个用户分类模型研究成果，可以为下一步针对不同用户群体进行需求侧响应方面的分析和预测研究打下基础。

本案例中用电行为分析主要包括数据采集、数据集成和存储、数据预处理和模型构建以及模型应用和预测结果展现，具体如图 10-6 所示。

具体用电行为分析中涉及的 4 个内容如下所示。

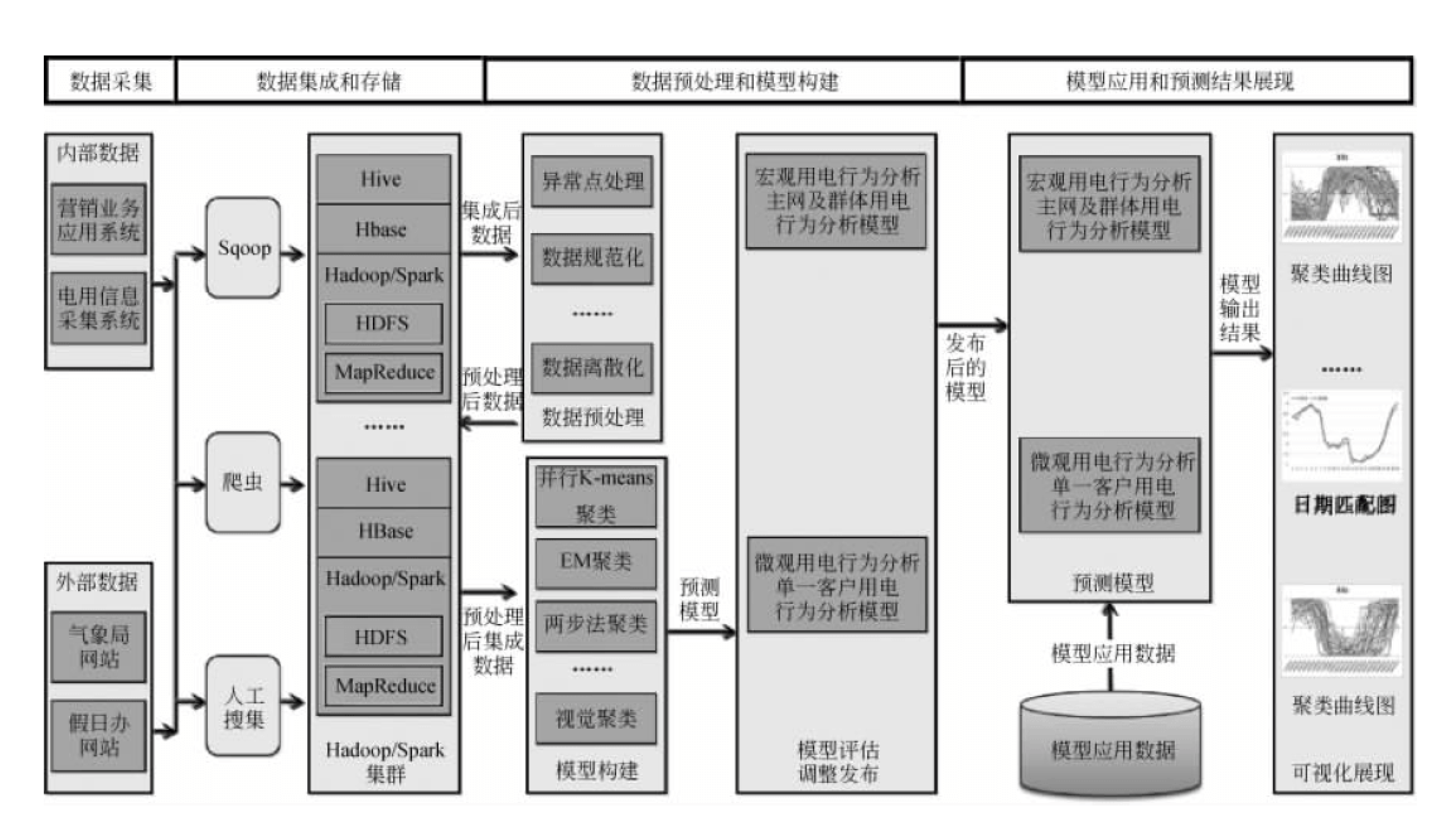


图 10-6　用电行为分析的总体框架图

1. 数据采集

数据采集是收集电网营销内部和外部的各类型、规模较大的数据，内部数据采集侧重在整合不同业务系统之间的数据，外部数据采集侧重在通过网络爬虫获取第三方数据。

1）内部数据采集

内部数据采集主要通过内部已有的、彼此分离的营销业务应用系统和用电信息采集系统等传统业务系统中提取数据。对于结构化的数据通过数据抽取转化工具来实现，包括数据的初始化、数据的增量抽取；对于非结构数据通过程序语言开发特定工具实现数据直接采集。

内部数据采集的主要困难在于打破部门职能藩篱、进行跨部门数据共享的思维和意识。只有充分认识到内部数据采集共享的价值，才能打破内部数据孤岛，实现数据的互联互通。内部数据的抽取有利于电网公司整合内部资源，一定程度上能加强数据的可见性、协作性与创造性，从而帮助电网企业大幅节约成本，提升竞争优势。但内部数据的维度相对集中，属于特定领域的点状数据。需要整合更多的外部数据来扩展数据的维度，丰富数据的属性刻画，帮助电网企业进行已有业务的优化和新增业务的衍生。

2）外部数据采集

外部数据采集一般有三种实现方式：第一种方式可以通过合作伙伴进行数据共享，第二种方式可以通过数据交换实现第三方公司的数据采集，第三种方式可以通过网络爬虫获取外部数据。第四种方式可以通过手动的方式进行外部数据处理。

这里重点阐述第三种数据采集方式：网络信息抓取服务。网络数据爬虫是指输入固定的数据源，按照一定的过滤和数据获取规则，自动抓取互联网网上数据。网络数据爬虫的难点有两个，一个是数据源网站的反爬虫设置，如通过用户请求的 Headers、基于用户行为反爬虫等技术，造成的数据抓取困难；一个是数据源本身以图片为载体造成的数据获取困难。

针对第一个难点，需要针对源网站的反爬虫技术，分别采用相应的破解技术，如在爬虫中修改或者添加 Headers、使用 IP 代理等方式，实现相关数据的顺利爬取。针对第二个难点，可以利用特定技术，如光学字符识别方法，通过图像增强、锐化、边缘检测等方法，进行特征提取和模型训练，从而获取图片中的数据。

本案例中天气和节假日等外部数据的爬虫有利于电网公司整合外部数据资源，使得点状数据扩展为带状和网状的数据，更好地增强数据的互联性、智能性、弹性和快捷性，激发出全新的工作思路和工作方式，利于已有业务的财务表现和新增业务的价值构建。

2. 数据集成和存储

对这些数据进行存储和处理，基于电网营销数据量越大，越需要各种不同种类的存储，且电网营销数据运营的大数据架构是可扩展的。

1）数据存储

数据存储主要提供分布式的存储功能。数据存储面临的主要挑战是数据大容量通常可达到 PB 级的数据规模，那么对于海量数据存储系统扩展能力的要求也会很高。同时，这海量的数据中存在大量信息是无效的，有效信息可能只分布在一个较短的时间段内，大量的数据存储给数据库带来不小的压力，而无效的数据更是对于资源的浪费。

存储层一般由 Hadoop 生态系统以及关系型数据库构建，Hadoop 生态系统主要用于海量数据的存储，主要包括 HDFS 分布式文件系统、HBase 分布式数据库以及 Hive 分布式数据仓库。存储层支撑对 PB 级别甚至更大级别的数据整合、存储与管理，同时具有强大的容错能力和平滑的线性扩展性。HDFS 分布式文件系统是整个大数据存储的基础，在它之上构建 HBase 分布式的数据库和 Hive 数据仓库。

2）数据计算

数据计算主要提供分布式的计算与处理功能。数据计算的挑战是数据处理速度响应的及时性。系统中的分布式数据计算模型一般基于 Spark 框架来实现，应用于大数据的实时处理和批处理计算分析任务执行，可以解决大数据计算中的交互查询及流式计算等核心问题。

Spark 由于提供了一套支持 DAG 图的分布式并行计算的编程框架减少多次计算之间中间结果写到 HDFS 的开销，提供 Cache 机制来支持需要反复迭代计算或者多次数据共享减少数据读取的 IO 开销，使用多线程池模型来减少 task 启动开销等诸多优点，是满足电网业务数据计算需求的最佳选择。

3. 数据预处理和模型构建

利用数据挖掘技术，在传统结构化数据处理上为用户提供越来越多的实时挖掘和分析，这些实时的智能服务可以支持实时的决策制定。

数据预处理主要是对质量较低的数据，如缺失、不准确、异常的数据，进行缺失值填充、异常值检测、离群值检测和规范化处理等。这里对缺失值填充进行侧重说明。缺失值是指在数据采集与整理过程中丢失的内容。缺失值的处理一般有两种方式，一是删除对应的记录，例如在用电行为分析中，如果某客户某一天没有记录，出现缺失，则将该客户当天所有负荷信息全部从数据库中删掉。这种方式在数据缺失非常少的情况下是可行的，但如果各个项目中都有少数的数据缺失，对所有缺失的记录都进行删除可能就会使总样本量变得非常

小，从而损失许多有用信息。缺失值处理的第二种方式是进行插值处理，所谓插值，是指人为地用一个数值去替代缺失的数值。在用电行为分析中，为保证数据分析的数量和准确性，一般进行第二种插值处理，且采用“均值”和“众数”两种方式进行填充。

数据挖掘和分析是从大量的数据中通过算法搜索信息并发现隐藏于数据中有趣、有用的模式和关系的过程。数据挖掘和分析的主要难点是如何将技术和算法应用到实际业务中去，从而提升传统的电网营销业务运作效率并改变运营模式。

本案例中主要涉及的数据挖掘和分析模型算法是聚类。聚类是把一组未带类别标签的数据分群，使“类内距离最小，类间距离最大”，包括并行 K-means、EM、两步聚类和视觉聚类等。电网公司进行用电行为分析，确定客户群体的类别、客户用电行业背景分析、客户电力衍生品购买趋势预测、大客户用电市场的细分等。

4. 模型应用和预测结果展现

模型应用是基于电网营销的数据和外部数据的模型分析结果通过可视化的手段进行展示，为领导层提供管理经营决策支撑。

可视化是利用计算机图形学和图像处理技术，涉及计算机图形学、图像处理、计算机视觉、计算机辅助设计等多个领域，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。本案例主要是聚类的模型结果展示。

10.3.3 宏观层面用电行为分析

宏观层面的用电行为分析方案主要包括下面三个内容。

(1) 历史主网负荷分群。主要基于 EM 聚类实现某区域主网的负荷分群，完成不同日期下的全省负荷特点。

(2) 二次聚类。基于主网的历史负荷分群结果，进一步分析这些日期群下非居民/居民的负荷特点。按照曲线相似性度量以及最大负荷值，对比这些用户的负荷曲线和主网同期的负荷曲线，可以得出“迎峰型”“逆峰型”等用电特征类型。

(3) 待分析日的历史相似日匹配。按照待分析日和历史日的最佳匹配关系，得到待分析日的“移峰填谷”模式匹配结果。

具体如下。

1. 历史主网负荷分群

从某区域的调度信息化系统中，获取某区域的主网历史一年多的主网负荷数据，采样时间范围为 2015 年 01 月 01 日至 2016 年 05 月 10 日，采样间隔为 15min，每个用户每天采样 96 点数据，见表 10-3。

表 10-3 EM 聚类的输入

日期	t_1 时刻负荷	t_2 时刻负荷	…	t_{96} 时刻负荷
day1	370	410	…	390
day2	378	385	…	400
…	…	…	…	…
day n	405	395	…	410

根据最大期望 EM 算法的运行原理，主要是交替使用最大期望判别所属分布和极大似然估计概率参数这两个步骤，逐步改进模型的参数，使参数和训练样本的似然概率逐渐增大，最后终止于一个极大点，从而将主网负荷所有日期分成三个群体：聚类 0、聚类 1 和聚类 2，见图 10-7。

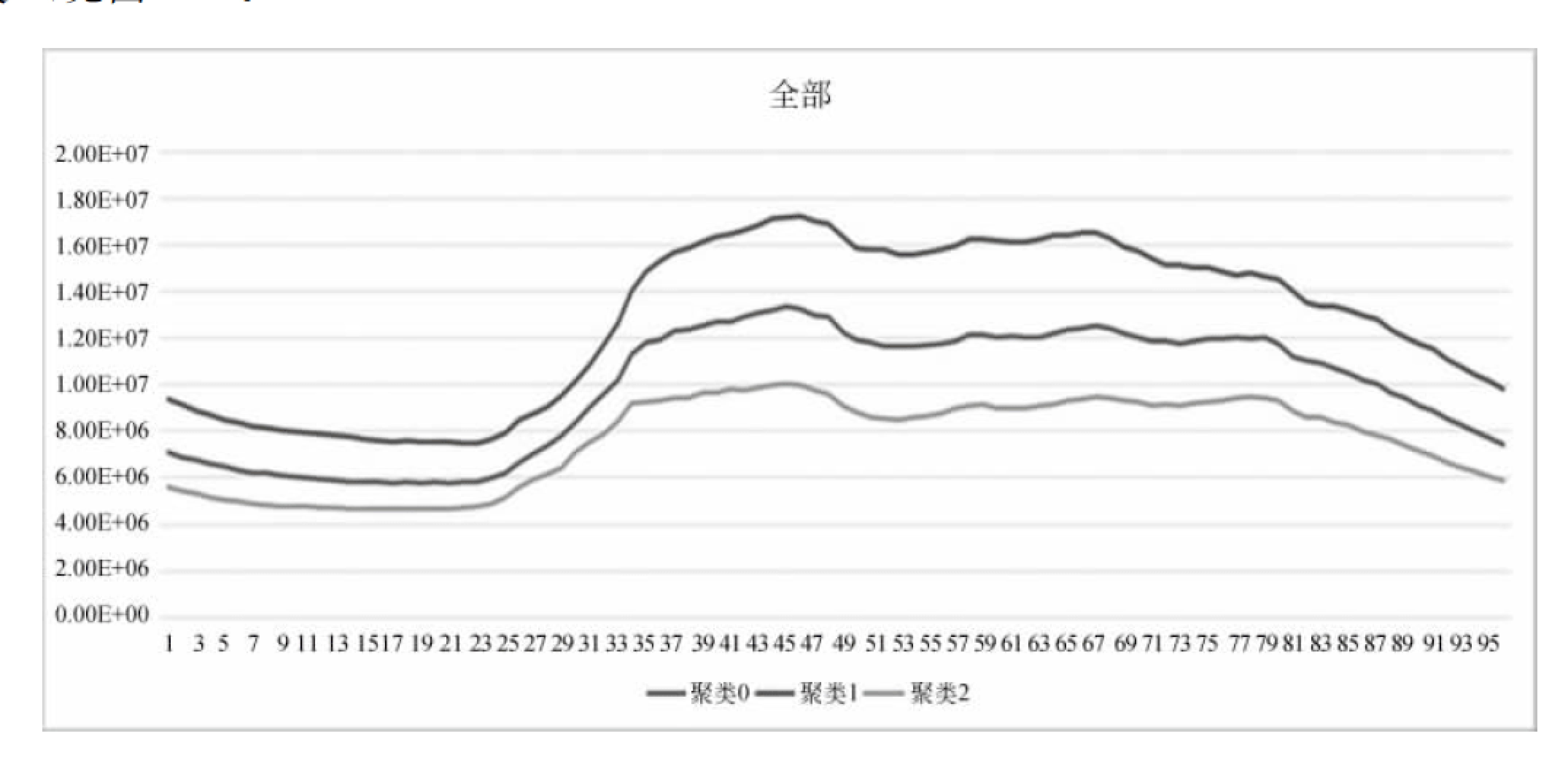


图 10-7　某区域主网聚类的结果

通过图 10-7 可以发现，聚类 2 的中心点负荷水平最低，聚类 1 的中心点负荷水平最高，聚类 0 的中心点负荷水平处于二者之间。结合这三个类的标号下不同日期的节假日分布和季节分布性质，可以得到各类主网的负荷行为特点，如表 10-4 所示。

表 10-4　主网聚类结果的特征分析

类名	类别	节假日	周末	工作日	春	夏	秋	冬
普通型	聚类 0	18.52%	34.09%	33.60%	46.67%	31.87%	29.35%	22.83%
高峰型	聚类 1	0.00%	30.68%	32.00%	20.00%	0.00%	65.22%	31.52%
假日型	聚类 2	81.48%	35.23%	34.40%	33.33%	68.13%	5.43%	45.65%

可以看出，“高峰型-聚类 1”的负荷都比较高，几乎不包含节假日，这些日期主网的负荷水平处于较高位置，调峰压力较高。“假日型-聚类 2”的负荷相对最低，主要是节假日期间对主网负荷贡献率较高的工业型负荷处于休息状态。而“普通型-聚类 0”的负荷处于两种之间的水平。

2. 二次聚类

进一步地，针对聚类 0、聚类 1 和聚类 2 分别进行群体用户的用电特征分群任务。这里收集了该区域 2015 年 01 月 01 日至 2016 年 05 月 10 日共 535 个用户用电数据，数据采集频率为每隔 15min 一次，每个用户每日采取 96 个点负荷数据。

在聚类前需要对用户的负荷数据进行预处理，主要是缺失值和异常值的处理。缺失值采用近一周的负荷平均值填充，对于异常值采取“3-西格玛法”。异常值检测和处理具体如下述方法。

一是纵向判断。设某用户 i 天 t 时刻的负荷为 $y(i,t)$，检验历史不同日期 n 天下该测量

点、该时刻的负荷平均值为 $\bar{\mu}_t(1,\cdots,n)$，标准差为 $\delta_t(1,\cdots,n)$，若 $|y(i,t)-\bar{\mu}_t(1,\cdots,n)|>3\times\delta_t(1,\cdots,n)$，则该点在纵向上为异常值。

二是横向判断。设某测量点 i 天 t 时刻的负荷为 $y(i,t)$，检验 i 天下不同时刻的 m 个负荷平均值为 $\bar{\mu}_i(1,\cdots,m)$，标准差为 $\delta_i(1,\cdots,m)$，若 $|y(i,t)-\bar{\mu}_i(1,\cdots,m)|>3\times\delta_i(1,\cdots,m)$，则该点在横向上为异常值。

如以上两条全部满足，这个点对应的数值为异常值，需要进行替换，替换时用纵向的平均值进行替换：

$$\bar{y}(i,t)=\begin{cases}\bar{\mu}_t(1,\cdots,n)+3\times\delta_t(1,\cdots,n),y(i,t)>\bar{\mu}_t(1,\cdots,n)\\ \bar{\mu}_t(1,\cdots,n)-3\times\delta_t(1,\cdots,n),y(i,t)<\bar{\mu}_t(1,\cdots,n)\end{cases}$$

其中，该点异常值替换后为 $\bar{y}(i,t)$；最后检验：若 $\bar{y}(i,t)$ 小于 0，则置为 0。

数据预处理结束后，分别对聚类 0、聚类 1 和聚类 2 这些日期下的用户群进行二次聚类，过程主要的方法是利用并行 K-means 算法完成二次聚类，主要包括下述三个步骤。

初始化步骤：首先确定聚类个数，随机确定各类中心，得到初始的<类标号，属性值>，同时将原始数据集分成若干个数据块。

Map 步骤：将每一个数据子集对应分配给一个 map 函数，针对每个数据块，计算每个数据项所属类别。类标号是当前样本相对于输入数据文件起始点的偏移量，map 函数首先将属性值解析成当前样本的各个维度的坐标值，然后基于欧式距离公式计算其与 k 个中心点的距离，找出与该样本最近距离簇的下标，即每个数据点都被匹配到对应的<类标号，属性值>。

Reduce 步骤：归并各块聚类结果得到完整的聚类结果，重新计算类中心作为下一次迭代的输入。将 map 阶段分群结果进行集合，新集合中所有对应键的相同值被归类在一起。reduce 函数首先解析出从层级合并中处理的样本个数和相应节点各个维度累加的坐标值，然后将对应值分别相加，除以总样本个数，即获得新的中心点坐标，形成一个输出的键/值对<类标号，属性值>，继续进行下一次迭代直至算法收敛。

基于上述三个步骤得到聚类 0、聚类 1 和聚类 2 的二次聚类下用户用电规律曲线。

结合二次聚类后中心点最大负荷和由皮尔森系数计算得到的相关系数，二次聚类后曲线最大负荷越大，对主网的负荷影响力就越高，是进行“是否进行负荷控制”时曲线匹配的最重要的影响因素；皮尔森系数大于 0.7 说明该用户负荷和主网负荷曲线相似度高，小于 0.7 说明二者相似度低，而皮尔森系数为负的时候说明该用户负荷和主网负荷曲线相似度为负向相关。

观察曲线和计算结果，可以发现：聚类 0 的用户的用电规律可以归纳为三种，在同一日期和时刻下，对于“普通型-聚类 0”主网负荷，主网负荷高的时候第一子类曲线所对应的这些测量点对应的负荷也比较高，二者相关系数为 0.9863，属于“削峰填谷”对象，如图 10-8 所示。

聚类 1 的用户的用电规律可以归纳为 4 种，在同一日期和时刻下，对于“高峰型-聚类 1”主网负荷，主网负荷高的时候第一子类曲线所对应的这些测量点对应的负荷也比较高，二者相关系数为 0.9799，属于“削峰填谷”对象，如图 10-9 所示。

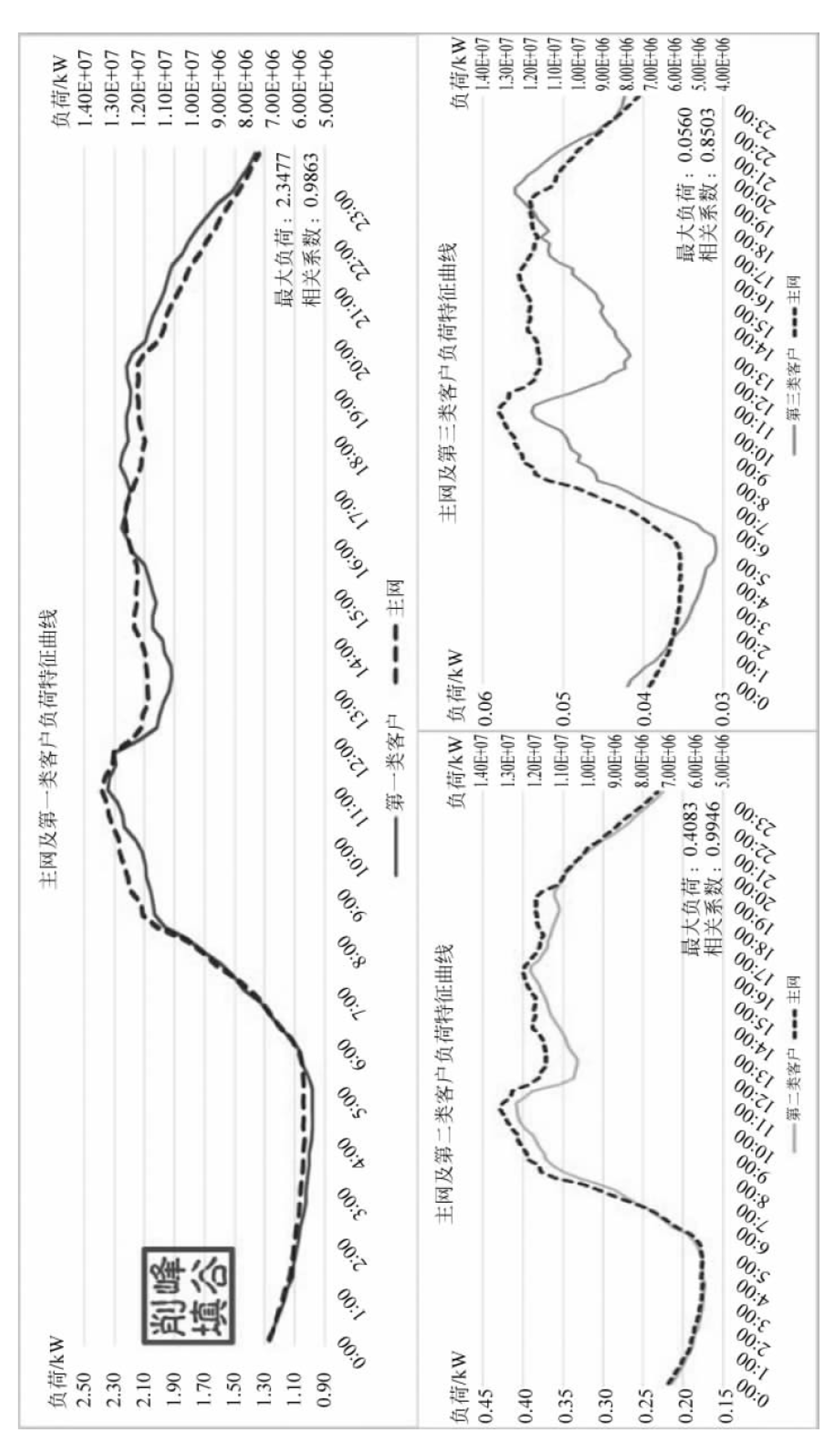


图 10-8 某区域主网聚类 0 的二次聚类结果

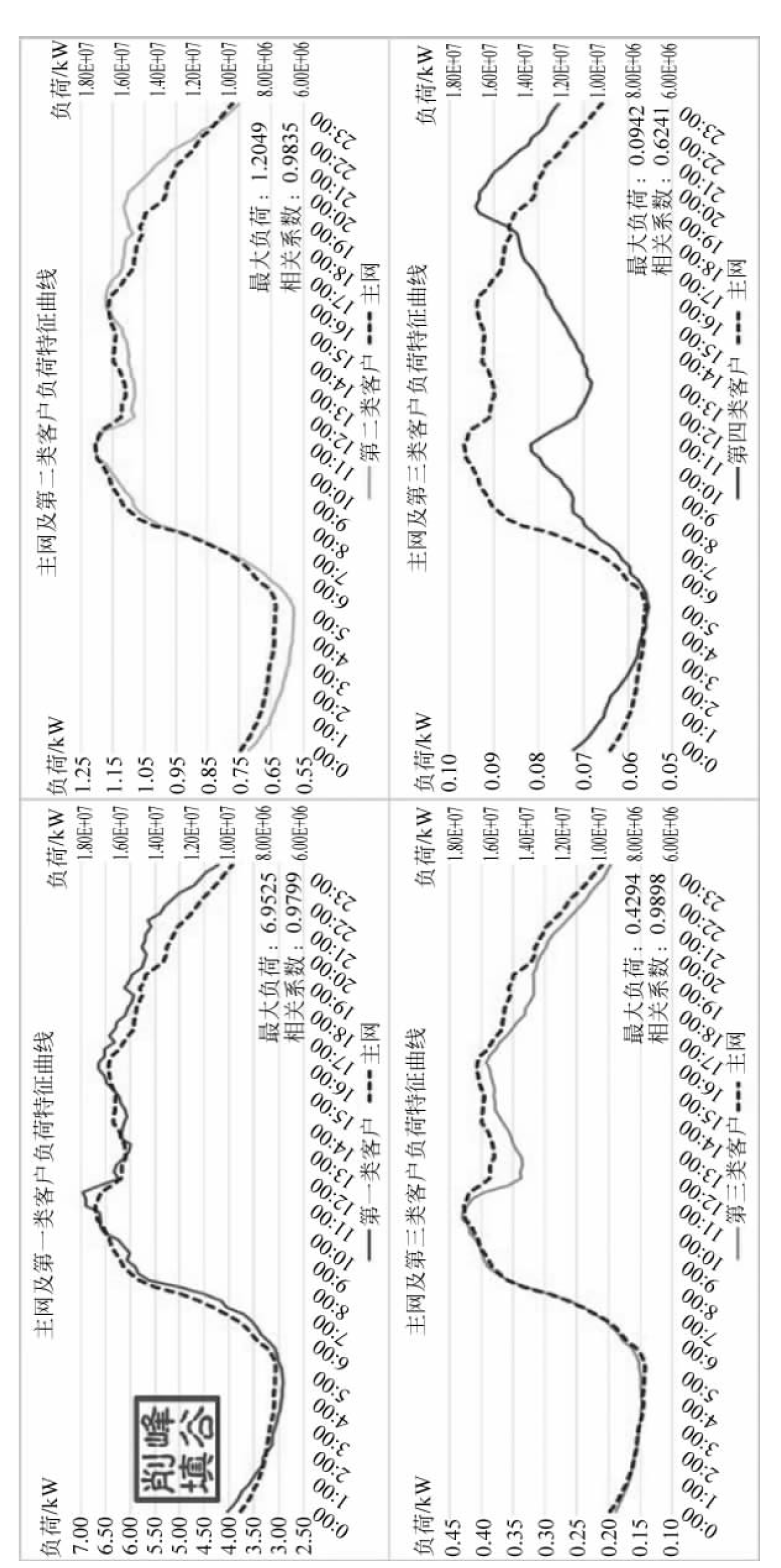


图 10-9 某区域主网聚类 1 的二次聚类结果

聚类 2 的用户的用电规律可以归纳为 5 种，在同一日期和时刻下，对于“节假日型-聚类 2”主网负荷，主网负荷高的时候第一子类曲线所对应的这些测量点对应的负荷也比较高，二者相关系数为 0.9764，属于“削峰填谷”对象；主网负荷高的时候第四子类曲线所对应的这些测量点对应的负荷比较低，二者相关系数为 −0.5998，属于“鼓励用电”对象；如图 10-10 所示。

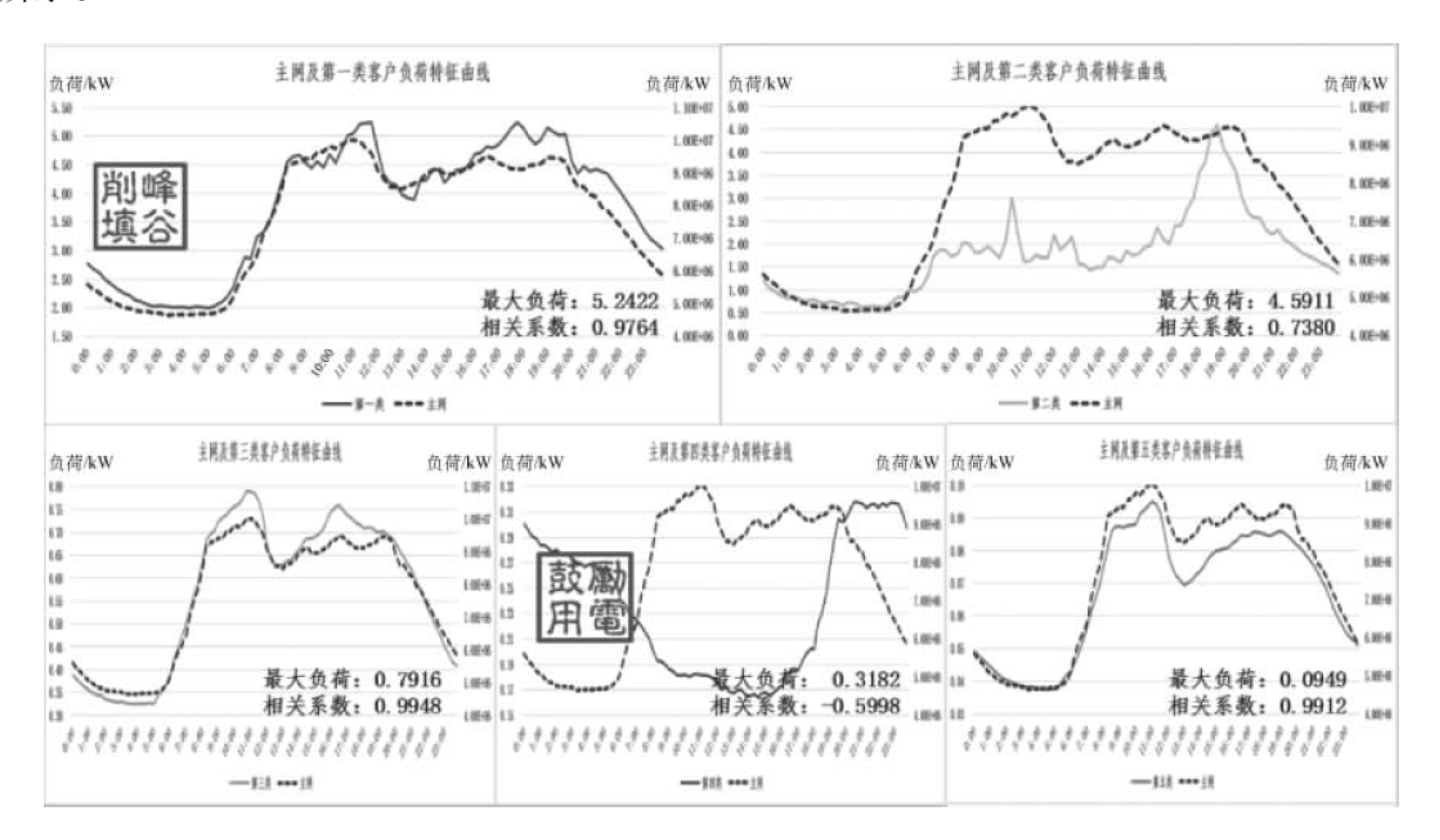


图 10-10　某区域主网聚类 2 的二次聚类结果

3. 待分析日与历史日匹配

在当前年度应用二次聚类结果时，根据历史用户负荷特征的错峰用电策略，利用动态时间规整算法找到当前年度待分析日和历史日的相似匹配，即可以得到当前年度待分析日的潜在的用电行为。

将为期一年的日期划分为三个日期集合：节假日、周末以及工作日。当待测日为节假日时，直接用历史节假日所归属的群体进行用电行为分析。当待测日为周末或者是工作日时，需要根据温度利用动态时间规整算法分别在历史周末集合和工作日集合中寻找相似日。待分析日和目标日匹配过程如图 10-11 所示。

动态时间规整算法的输入为历史年时间段内周末或工作日温度序列 $T_{old}=(T_1,\cdots,T_m)$，当前年同时间段内且包含待测日的周末或工作日温度序列 $T_{new}=(\overline{T}_1,\cdots,\overline{T}_n)$，满足 $m\geqslant n$。实现过程是搜索从 $(1,1)$ 点出发搜索至 (m,n)，可以展开若干条路径，可计算每条路径达到 (m,n) 点时的总的积累距离，通过逐点向前寻找就可以求得整条路径，具有最小累积距离者即为规整路径。动态时间规整算法的输出为最小规整距离，以及历史日期点和当前日期点的匹配关系结果。

这样，利用动态时间规整算法，每个待测日都可以找到历史中的相似匹配日，按照相似日所对应的主网负荷表现，若此客户取避开高峰用电方式，有利于电网安全运行，不是开展错避峰用电措施的重点。

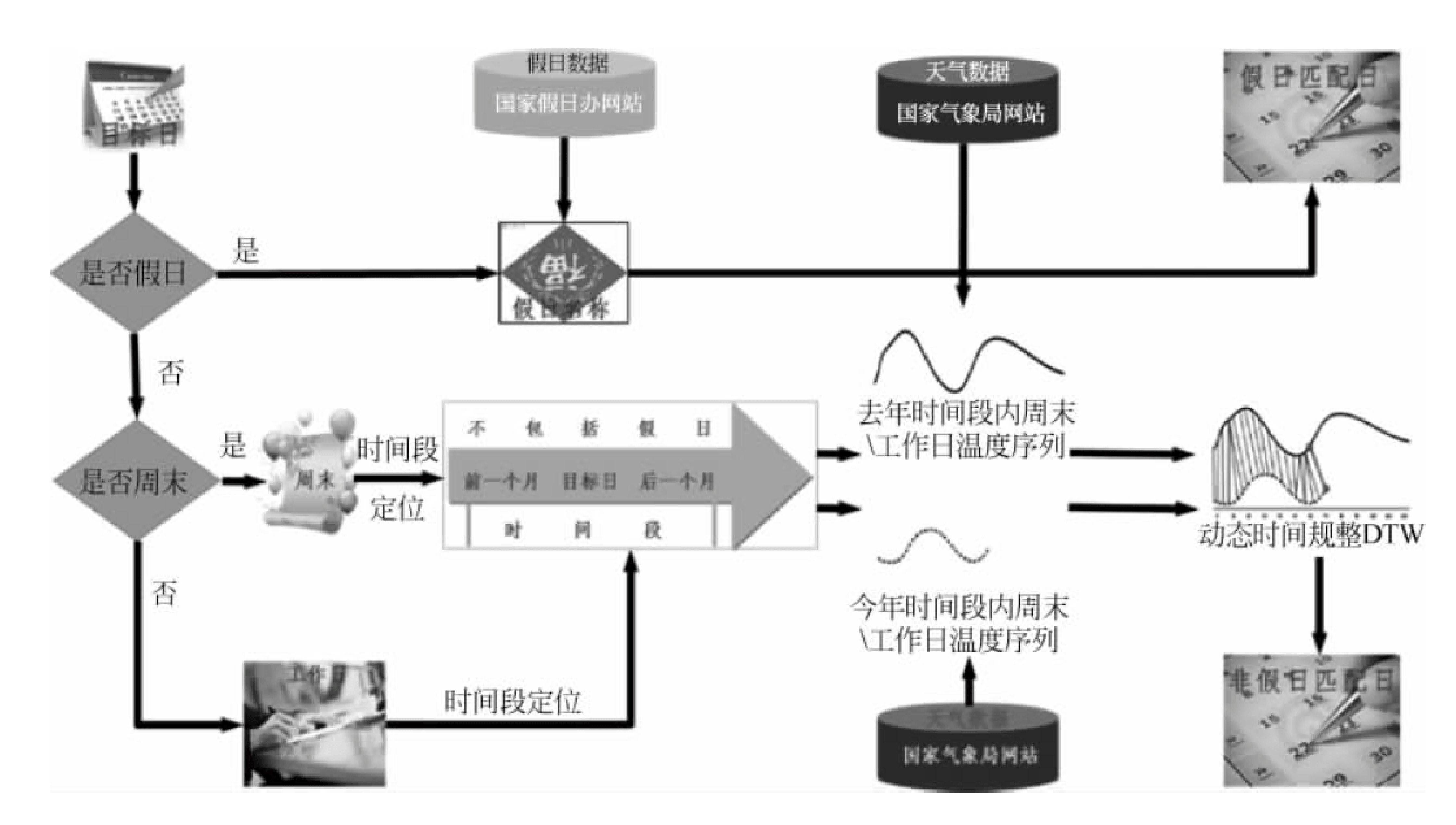


图 10-11 待分析日和目标日匹配过程

如表 10-5 所示为一个测量点日期匹配示例。

表 10-5 测量点日期匹配示例

用户 ID	当前年待分析日	历史年的相似日	聚 类 标 号
用户 A	2016-05-08	2015-04-25	主网聚类 1 的第一子类
用户 B	2016-05-08	2015-04-25	主网聚类 2 的第四子类

通过分析每类客户用电负荷特征，识别具有移峰填谷潜力客户，并提出负荷管理建议。针对待分析日计算得到匹配历史日后，若负荷曲线特征和主网曲线特征高度相似，且自身负荷曲线最大负荷值比较高，对电网安全运行影响较大，是开展错避峰措施的重点客户，同时需要关注该类客户电气设备安全使用情况。

针对这类“迎峰型”重点客户（如表 10-5 中的用户 A），建议开展可中断负荷及补偿分析工作，分析重点客户移峰填谷潜力，在错峰用电时需要给予重点关注和有序用电引导。具体实施时，业务人员可以结合业务经验和用户档案信息，进一步找到更精准的“有序用电”用户，为电网平稳、安全运行提供辅助决策支撑。如业务人员分析用户 A 的客户档案信息筛选出其为“单班次生产”“高耗能-造纸”的工业用户，则可作为实际操作时的错峰用电对象进行引导。

而用户负荷曲线特征均属于“逆峰型”的（如表 10-5 中的用户 B），已基本采取避开高峰用电方式，对主网负荷总体影响不大，有利于电网安全运行，不是开展错避峰措施的重点对象，在错峰用电时需要给予鼓励保持用电习惯。

10.3.4 微观层面用电行为分析

微观层面的用电行为分析方案主要包括下面三个内容。

（1）单个居民型用电行为分析。将用电类型为城镇居民和农村居民的用户筛选出来，

针对每一个居民用户，利用 EM 算法按照不同日期下的负荷数据进行分群，掌握单一居民用户的用电特征。

（2）单个工业型用电行为分析。将用电类型为一般工业和大工业的用户筛选出来，针对每一个工业型用户，利用两步聚类算法按照不同日期下的负荷数据进行分群，掌握单一工业用户的用电特征。

（3）单个商业型用电行为分析。将用电类型为一般工商业的用户筛选出来，针对每一个商业用户，利用视觉聚类算法按照不同日期下的负荷数据进行分群，掌握单一商业用户的用电特征。

1. 单个居民型

本文共收集了 160 户居民家庭用电数据，采样时间范围为 2015-01-01～2016-05-10，采样间隔为 15min，每户居民每天采样 96 点数据。将这些数据去除噪声（有些采样值很大）后全部作为单个居民型用电行为分析的实验数据，以此为基础对居民用户类型的用电特征展开研究。

单一客户负荷分群的目的是根据客户各种负荷数据，将具有相同特征的日期聚集在一起，不同日期的负荷分散开，因此这里采用 EM 聚类算法进行分群，符合单一客户不同时间下负荷特征分群的要求。

EM 聚类结束后，根据自定义的聚类性能评价指标进行模型的效果评估。评价指标自动判断聚类个数，利用算法自动筛选最优化的聚类个数。评价指标设计的原则是“类间距离最大，类内距离最小”，具体如下：

$$\mathrm{POC}=\left(\left(\frac{\sum_{i=1}^{k}\sum_{j<i}^{k}\|V_i-V_j\|^2}{C_k^2}\right)\Bigg/\left(\frac{\sum_{i=1}^{k}\frac{\sum_{j=1}^{n}\|x_j-V_i\|^2}{n}}{k}\right)\right)\Bigg/k$$

其中，k 表示类数，n 表示对于第 i 类有 n 个样本，V_i表示第 i 类的类中心，x_j表示第 i 类第 j 个样本值；k 在一个指定区间内（例如 $k\in[1,5]$），求 POC 值最大，代表该类为最佳类。

按照上述方法，可以得到单一居民在不同日期下的负荷的不同特性。对这些聚类结果的解读，一方面可以指导居民更好地用电，另一方面可以让电网企业了解居民用电的特点，为制定分时电价时提供支撑。

下面通过一个居民型用户的结果示例进行说明：通过 EM 聚类和最佳聚类评估指标，将该用户一年内的日期负荷分为两个群体，见图 10-12。

进一步地，联系第一类和第二类标号所对应的日期的季节分布特点，很明显地发现：第一类负荷曲线的平均负荷较高，时间分布以夏季和冬季为主；第二类负荷曲线的平均负荷较低，以春季为主，如表 10-6 所示。

由此，联系该居民用户所属的区域，该区域为典型的中国内陆南方城市，夏季温度较高，需要空调等进行室内制冷；冬季温度降低，需要空调等进行室内升温。

同时，观察一天的曲线平均负荷，可以明显观察到三个小高峰，早晨 7 点至 8 点，中午 11 点至 12 点，下午 17 点至 18 点，如图 10-13 所示。结合居民用电的特点，可以推测为早餐、中餐和晚餐的时间，因为一些咖啡机、面包机、电饭煲和烤箱等家用电器的阶段性工作引起该时段的负荷出现波动。

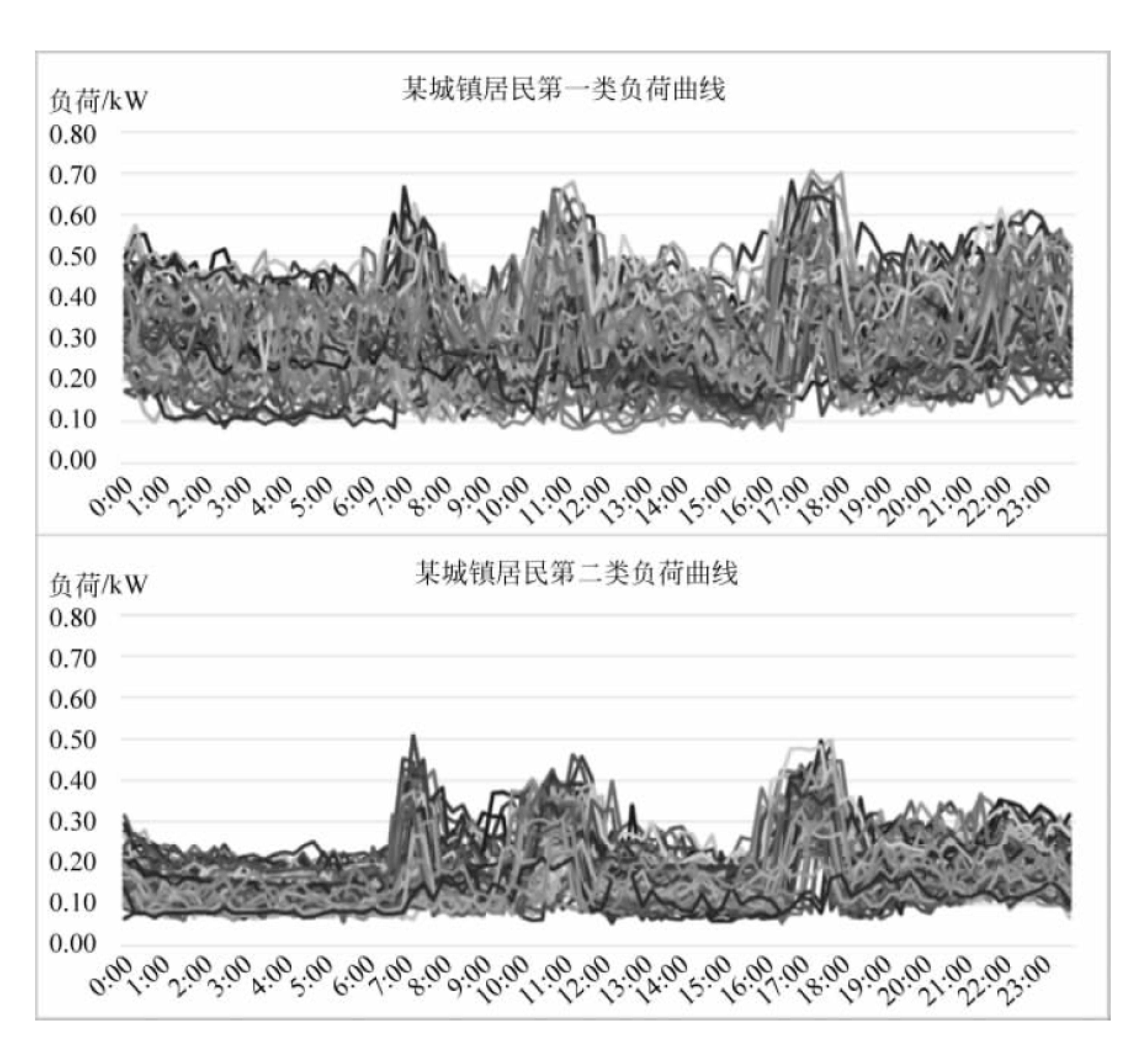


图 10-12　某城镇居民聚类结果

表 10-6　某城镇居民聚类特征

	第　一　类				第　二　类			
季节	春季	夏季	秋季	冬季	春季	夏季	秋季	冬季
占比	26.09%	80.43%	53.85%	92.59%	73.91%	19.57%	46.15%	07.41%



图 10-13　某城镇居民聚类中心点负荷曲线

2．单个工业型

本文共收集了 215 户工业型用电数据，采样时间范围为 2015-01-01～2016-05-10，采样间隔为 15min，每户工业用户每天采样 96 点数据。将这些数据去除噪声（有些采样值很大）后全部作为单个工业型用电行为分析的实验数据，以此为基础对工业用户类型的用电特征展开研究。

考虑到工业型用户（尤其是大客户）数量大，且对整个电网企业的盈利贡献占比较高，达到总体 80%以上，对这些工业型的负荷特征分析需要更慎重。因此在获取了这些用户的负荷数据后的第一步进行数据预处理，这里重点强调离群值检测。

离群值检测在预聚类步骤进行。将相对于其他子聚类具有较少记录的子聚类视为潜在离群值，且重新构建不包括这些记录的子聚类树。子聚类被视为包含潜在离群值的下限大小由百分比选项控制。如果其中某些潜在离群值记录与任何新子聚类配置足够相似，则可将其添加到重新构建的子聚类中。将其余无法合并的潜在离群值视为离群值添到“噪声”聚类中并排除在分层聚类步骤之外。

使用经过离群值处理的“两步”模型对数据进行评分时，会将与最近主要聚类的距离大于特定阈值距离（基于对数似然）的新观测值视为离群值分配到“噪声”聚类中，名称为－1。

在预处理后，开始利用聚类进行负荷特征分群。在聚类算法中，两步聚类可以非常迅速地对大量聚类解决方案进行分析并为训练数据选择最佳聚类数。通过设置最大聚类数和最小聚类数指定要尝试的聚类解决方案的范围。“两步聚类”通过一个两阶段过程确定最佳聚类数。在第一个阶段，随着所添加聚类的增多，可基于贝叶斯信息准则（BIC）中的差异选择模型中聚类数的上限。在第二个阶段，为聚类数比最小 BIC 解决方案还少的所有模型找出聚类间最小距离的差异。距离的最大差异用于识别最终聚类模型。

这样，利用两步法，可以得到单一工业用户在不同日期下的负荷不同特性。对这些聚类结果的解读，一方面可以指导工业用户开展有序用电，另一方面可以让电网企业了解工业用电的特点，在制定电价优惠时提供支撑。

下面通过一个工业型用户的结果示例进行说明：通过二步聚类，将该用户一年内的日期负荷分为两个群体，见图 10-14。

进一步地，联系第一类和第二类标号所对应的日期的季节分布特点，很明显地发现：第一类负荷曲线的平均负荷较低，第二类负荷曲线的平均负荷较高，但无论是第一类负荷还是第二类负荷，这两类负荷曲线在季节分布上差异不明显，如表 10-7 所示。

由此，联系该工业用户所属的行业类别为“谷物磨制”，该类型的企业在季节性上确实无太大差异，但一个统计周期内主要是以第二类负荷特征为主。

同时，观察一天的曲线平均负荷，可以明显观察到：第一类负荷曲线整体较低，工作节奏比较缓慢，包含两个小高峰：早晨 8 点至 12 点和下午 13 点至 17 点。第二类负荷曲线整体较高，工作时长比较大，包含三个小高峰：早晨 8 点至 12 点、晚上 18 点至 22 点。结合该工业用户的生产班次安排的特点，可以推测这家企业工作类型为“两班制”，需要在用电检查时多给予关注，以免发生违规不合格用电现象。

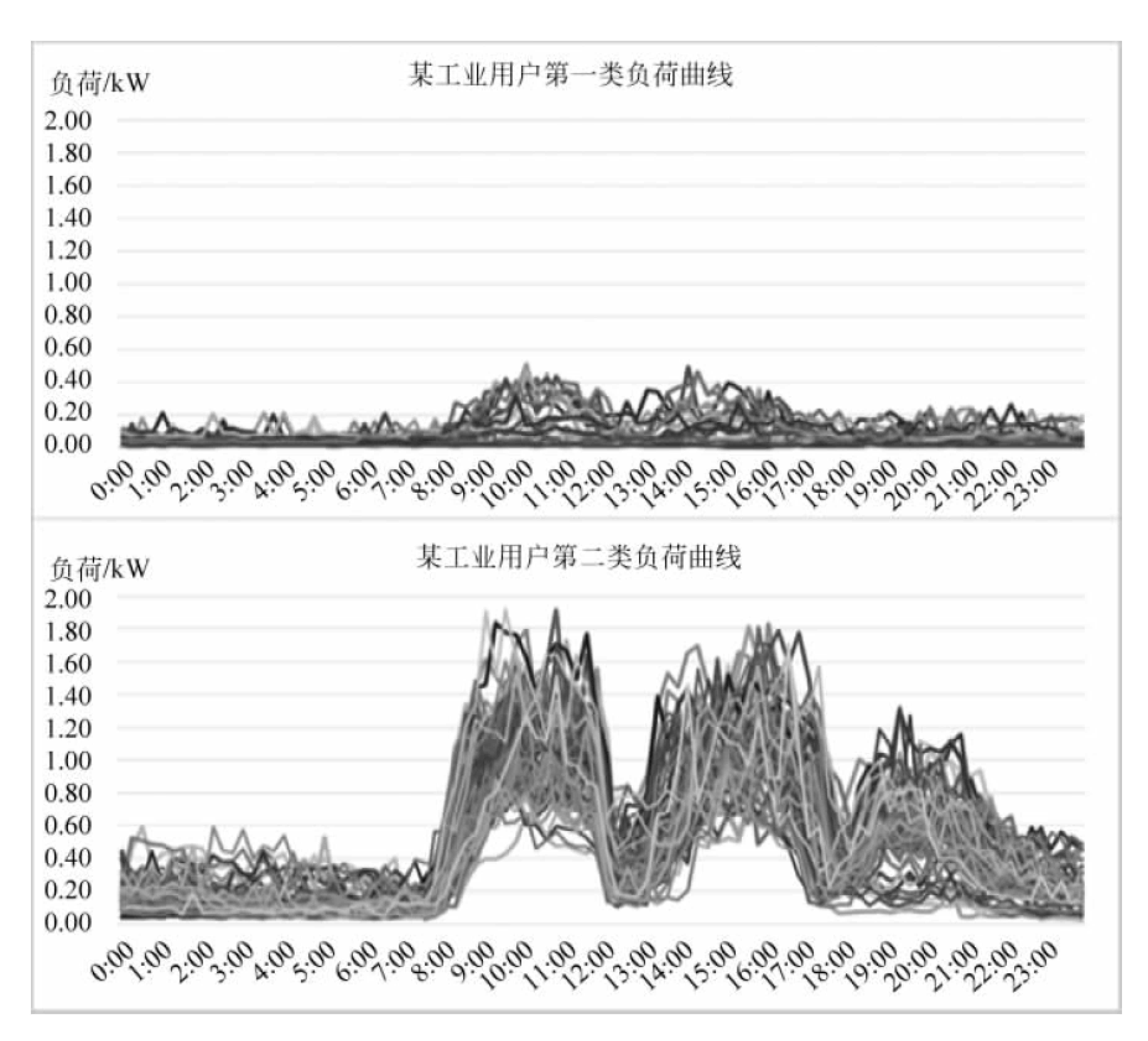


图 10-14　某工业用户聚类结果

表 10-7　某工业用户聚类特征

	第　一　类				第　二　类			
季节	春季	夏季	秋季	冬季	春季	夏季	秋季	冬季
占比	19.48%	23.53%	27.16%	27.27%	80.52%	76.47%	72.84%	72.73%

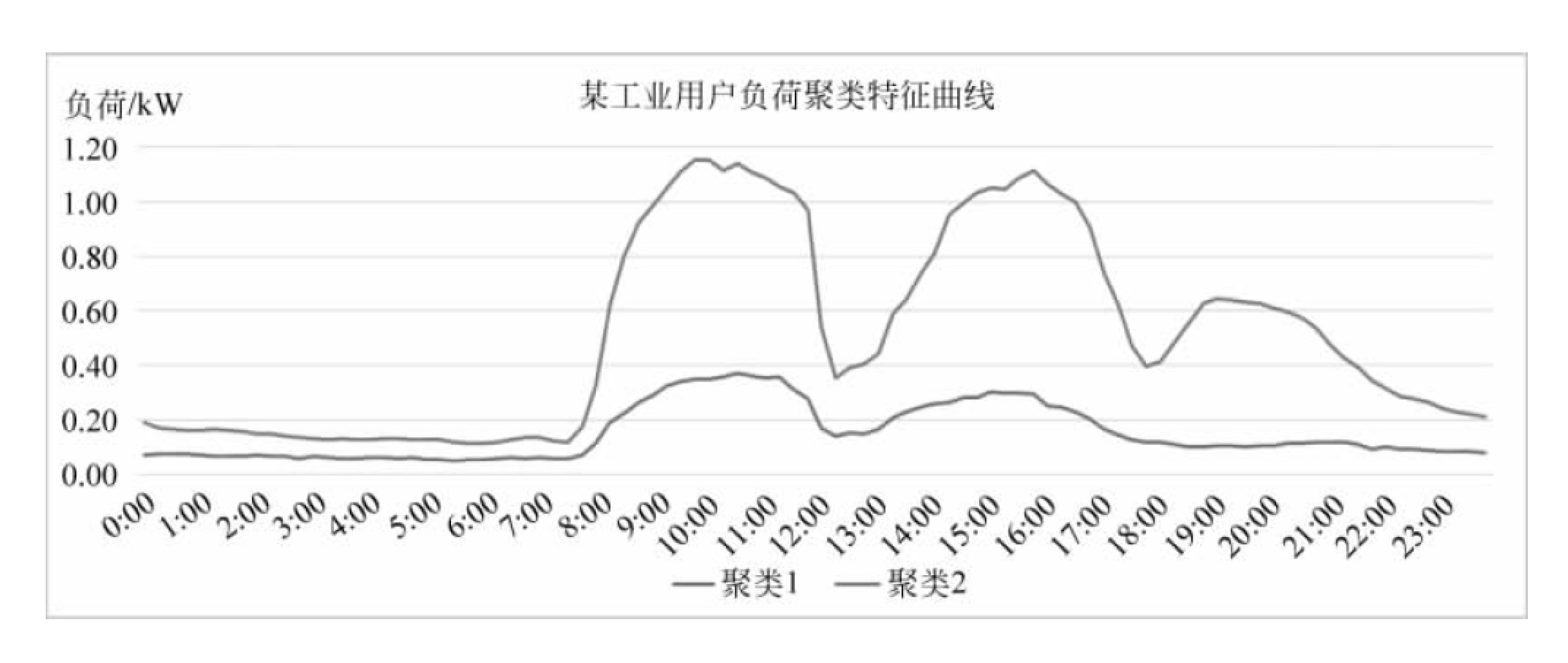


图 10-15　某工业用户聚类中心点负荷曲线

3. 单个商业型

本文共收集了 160 户商业用电数据，采样时间范围为 2015-01-01～2016-05-10，采样间

隔为 15min，每户商业用户每天采样 96 点数据。将这些数据去除噪声（有些采样值很大）后全部作为单个商业型用电行为分析的实验数据，以此为基础对商业用户类型的用电特征展开研究。

考虑到商业型用户对电网企业来说属于新增潜力比较大的群体，非常有必要单独将这部分用户进行特征分析。因此在获取了这些用户的负荷数据后，类似工业型用户，第一步需要进行数据预处理，这包括缺失值、异常值和离群值检测处理。

针对数据预处理之后的数据质量较高的数据，开始利用聚类进行负荷特征分群。在聚类算法中，视觉聚类可以较为准确地对实现分群并为训练数据自动选择最佳聚类数。在视觉聚类算法中，每一个数据点被视作空间的一个光点，于是一个数据集对应构成了空间中的一幅图像。当模糊化这一图像时，每一个小光点首先变化为一个小光斑。进一步地模糊，使得小光斑逐渐溶为大光斑。当分辨率充分低时，整个图形便成为一个光斑。如果将每个光斑看成一个类，则上述模糊化的过程便形成逐级聚类树，结点代表不同尺度聚类的类，父结点代表的类由子结点所代表的类融合而成。过程中尺寸变化维持时间最长的时候，对应的即为最佳聚类个数。

这样，利用视觉聚类法，可以得到单一商业用户在不同日期下的负荷不同特性。对这些聚类结果的解读，一方面可以指导商业用户开展有序用电，另一方面可以让电网企业了解商业用电的特点，在制定电价优惠时提供支撑。

下面通过一个商业型用户的结果示例进行说明：通过视觉聚类，将该用户一年内的日期负荷分为三个群体，见图 10-16。

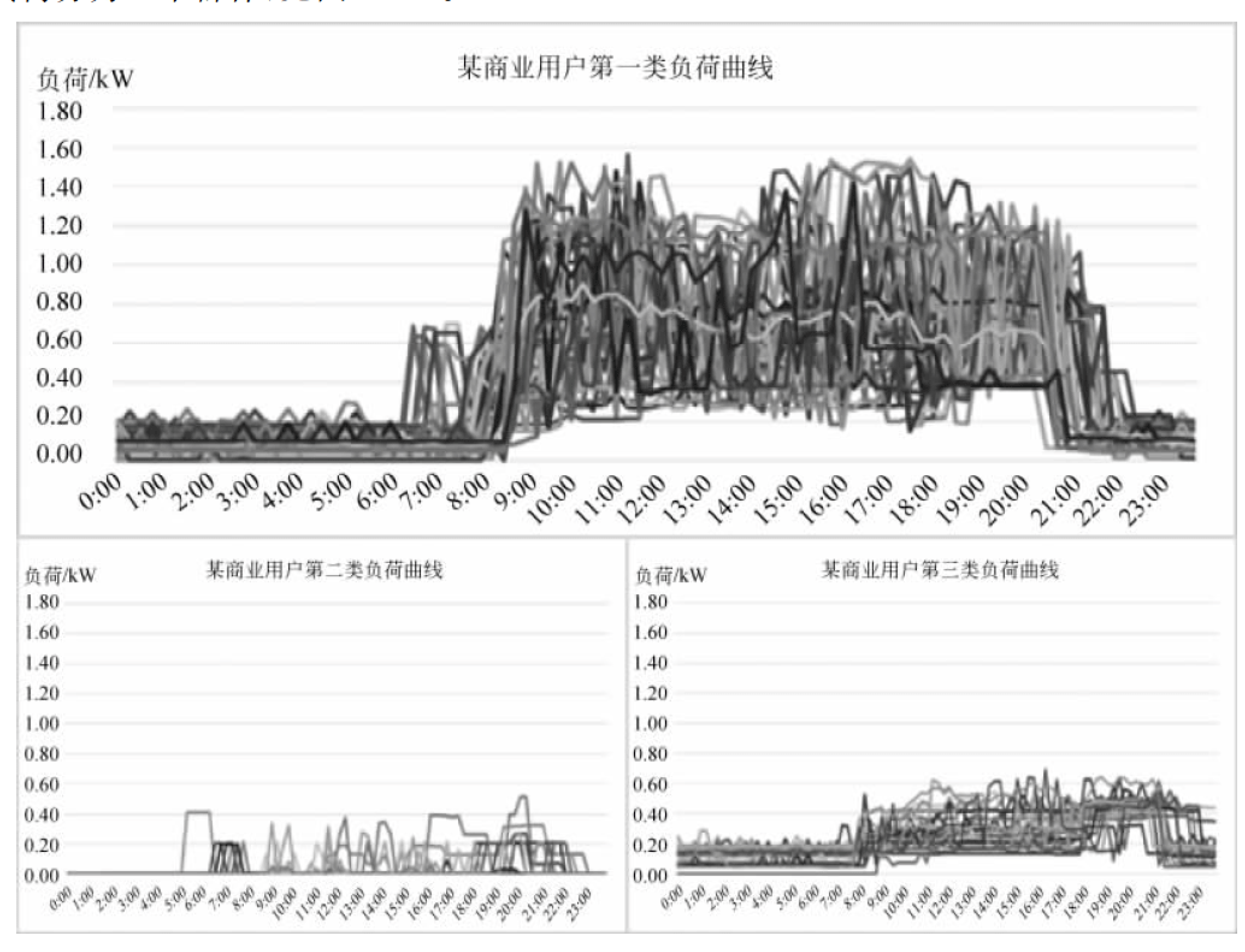


图 10-16　某商业用户聚类结果

进一步地，联系第一类、第二类标号和第三类所对应的日期的季节分布特点，很明显地发现：第一类负荷曲线的平均负荷最高，第二类负荷曲线的平均负荷一般水平，第三类负荷曲线的平均负荷水平最低。第一类负荷以夏季和冬季为主；第二类负荷以春季为主；第三类负荷以秋季为主，如表 10-8 所示。

表 10-8 某商业用户聚类特征

	第 一 类				第 二 类				第 三 类			
季节	春季	夏季	秋季	冬季	春季	夏季	秋季	冬季	春季	夏季	秋季	冬季
占比	07.61%	73.91%	18.68%	74.44%	50.00%	00.00%	00.00%	01.11%	42.39%	26.09%	81.32%	42.39%

由此，联系该商业用户所属的行业类别为“物业管理”，该类型的企业在每年年中，以及春节等时间上工作节奏强度最高，在季节上表现为夏季和冬季。春季属于新一年的规划时期，工作节奏也较高，相比较夏季和冬季，秋季工作忙碌程度上确实较低。

同时，观察一天的曲线平均负荷，可以明显观察到：第一类负荷曲线全天工作从早晨 8 点至晚上 22 点整体最高，中午和下午均无休息间断，一天的工作节奏比较紧张。第二类负荷曲线全天工作从早晨 8 点至晚上 22 点整体较高，尤其是下午 18 点至晚上 22 点出现小高峰，工作时长比较大。第三类负荷曲线全天基本接近为 0，排除掉停电故障后，可推测在这个群体下的日期里该商业用户的生产班次安排最为悠闲。



图 10-17 某商业用户聚类中心点负荷曲线

通过对微观层面的用电行为分析，完成了居民型、工业型和商业型用户的用电行为分析。通过横向对比，可以看出居民型、工业型和商业型用户三者之间在用电上差异较大。这样，在某区域的分类电价的制定时，相关的时间周期长度为电网企业确定平段、峰段和谷段等提供依据。在掌握用户负荷特征的基础上，可以更好地进行负荷的短期预测。此外，在进行需求侧管理时，通过对用户负荷特点进行把握，有利于未来微电网环境下小区域内的电力输配和调度供电管理。

综上，通过对用电行为分析的总体架构设计、宏观层面用电行为分析和微观层面用电行为分析，可以有效地将相关成果与电网企业现有的供电服务、有序用电、催缴回收、安全用电以及智能用电等业务场景结合起来，促进该区域电网企业对外客户提供更优质、可靠的供电服务。

伴随着智能电表等计量装置的普及，电网企业累积了越来越多的数据，数据分析的力量和价值将进一步地凸显，成为推动电网企业和用户互动参与的一股主要力量。大数据分析促进电网企业等公用事业机构利用创新的数据分析和数据架构方法解决企业运营中的问题，也为新业务领域商业化创造契机。

未来，电网企业通过更主动和科学地洞察电力用户的消费特点和趋势，为电力消费者提供更明智的电力产品、服务和解决方案，才能在变革的浪潮中勇立潮头。